MACHINE LEARNING APLICAT A LA PREDICCIÓ DEL RISC EPIDEMILÒGIC AL BRASIL

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Joaquim Bauxell Cornet

In partial fulfilment

of the requirements for the degree in

TELECOMMUNICATIONS TECHNOLOGIES AND SERVICES ENGINEERING

Advisor: Mercè Vall-llossera Ferran

Barcelona, January 2020

Abstract

This thesis consists in the development of predictive algorithms of the dengue epidemiological risk based on Machine Learning, using the climatological and socioeconomic variables that affect the spread of the disease throughout Brazil. Dengue is a virus that can be transmitted by the Aedes Aegypti mosquito and is common in subtropical areas.

This document details the process in which the aforementioned data has been collected, sorted and analysed. Also describing all the autonomous learning tools used for this purpose, and how the risk of dengue outbreaks can be predicted beforehand with certain conditions. Finally, the first results from the developed models are presented.

Resum

Aquest projecte consisteix en el desenvolupament d'algoritmes predictius del risc epidemiològic del dengue basats en Machine Learning a partir de les variables climatològiques i socioeconòmiques que afecten la seva propagació per tot el Brasil. El dengue és un virus que és transmès pel mosquit Aedes Aegypti, comú en regions subtropicals.

En aquest document es detalla el procés en el qual aquestes s'han recollit, ordenat i analitzat. També es descriuen totes les eines d'aprenentatge autònom que s'han fet servir per aquest propòsit, i com el risc de brots de dengue es pot predir amb unes determinades condicions. Finalment es presenten els primers resultats dels models desenvolupats.

Resumen

Este proyecto consiste en el desarrollo de algoritmos predictivos del riesgo epidemiológico del dengue basados en Machine Learning a partir de las variables climatológicas y socioeconómicas que afectan su propagación por todo el Brasil. El dengue es un virus que es transmitido por el mosquito Aedes Aegypti, común en las regiones subtropicales.

En este documento se detalla el proceso en el que estos datos se han recolectado, ordenado y analizado. También se describen todas las herramientas de aprendizaje autónomo que se han usado para este propósito, y como el riesgo de brotes de dengue puede predecirse bajo unas determinadas condiciones. Finalmente se presentan los primeros resultados de los modelos desarrollados.

Agraïments

M'agradaria expressar el meu agraïment a la Mercè Vall-llossera, per la oportunitat de poder realitzar aquest projecte i el seu guiatge. També a Hellen Gurgel, per compartir dades crucials. I finalment a la UPC per l'equip proporcionat.

Historial de revisions i aprovació

Revisió	Data	Propòsit
0	15/12/2020	Creació del document
1	12/01/2021	Revisió del document
2	23/01/2021	Revisió i aprovació del document

LLISTA DE DISTRIBUCIÓ DEL DOCUMENT

Nom	e-mail
Joaquim Bauxell Cornet	joaquim.bauxell@estudiantat.upc.edu
Mercè Vall-llossera Ferran	merce.vall-llossera@upc.edu

Escrit per:		Revisat i aprovat per:	
Data	15/12/2020	Data	23/01/2021
Nom	Joaquim Bauxell Cornet	Nom	Mercè Vall-llossera Ferran
Posició	Autor del projecte	Posició	Supervisora del projecte

Índex

Llista de figures

Llista de taules

1. Introducció

1.1. Objectius del projecte

El propòsit d'aquest projecte és crear un mètode viable per poder predir el risc epidemiològic del dengue al Brasil amb anterioritat i de manera fiable. S'ha utilitzat dades com ara la temperatura, la humitat del sòl, precipitacions, índex de desenvolupament municipal entre altres.

Per aconseguir-ho s'han seguit els següents passos:

1.- Trobar i modelar les dades que poden afectar la propagació del virus. S'han utilitzat dades dels anys 2010 al 2018.

2.- Determinar quines variables tenen un efecte major sobre el risc epidemiològic del dengue.

3.- Crear alguns algoritmes de Machine Learning per predir la situació epidemiològica amb un mes d'anticipació per tot Brasil.

4.- Visualitzar i interpretar els resultats.

1.2. Requeriments i especificacions

Els requeriments pel correcte desenvolupament d'aquest projecte són:

- Que les dades recollides tinguin una precisió suficient com per garantir l'anàlisi a nivell de cada municipi brasiler.
- Per a l'estudi, les dades recollides estan centrades entre els anys 2010 i 2018.
- Que la informació proporcionada per les dades recollides tingui una correlació suficient amb l'índex del risc de la propagació del virus per tal de poder garantir una bona precisió en la predicció del risc de propagació del virus.
- Les aplicacions de Machine Learning desenvolupades han de ser escalables. És a dir, es podrà dur a terme una predicció en qualsevol municipi o agrupació de municipis indistintament per tot el país.

Degut a la naturalesa del projecte les principals eines utilitzades són únicament de software ja que, per exemple, les dades de satèl·lit ja són d'accés públic. Aquestes eines utilitzades són:

- MATLAB: per a l'extracció de dades ambientals (NDVI, NDWI, temperatura i humitat) i les interpolacions i arranjaments conseqüents.
- Python en un entorn Anaconda: per la resta de tasques relacionades amb la programació.
- Pytorch en un entorn de Google Colab: per desenvolupar els diversos models de Machine Learning.

- Panoply: per la visualització íntegra dels diversos mapes ambientals.
- QGIS: per la visualització de la resta de mapes així com la confecció de mapes nous a partir de les dades i resultats obtinguts.
- PostgreSQL: a partir de pgAdmin. Utilitzat per la creació de la base de dades global del país, incloent la informació espacial amb PostGIS de cada un dels municipis, per així facilitar la visualització a posteriori sobre un mapa.

1.3. Pla de treball

La tasca ha estat dividida en cinc paquets de treball. En la següent figura es detalla el temps invertit en cadascun d'ells, així com en cadascuna de les tasques que els conformen.

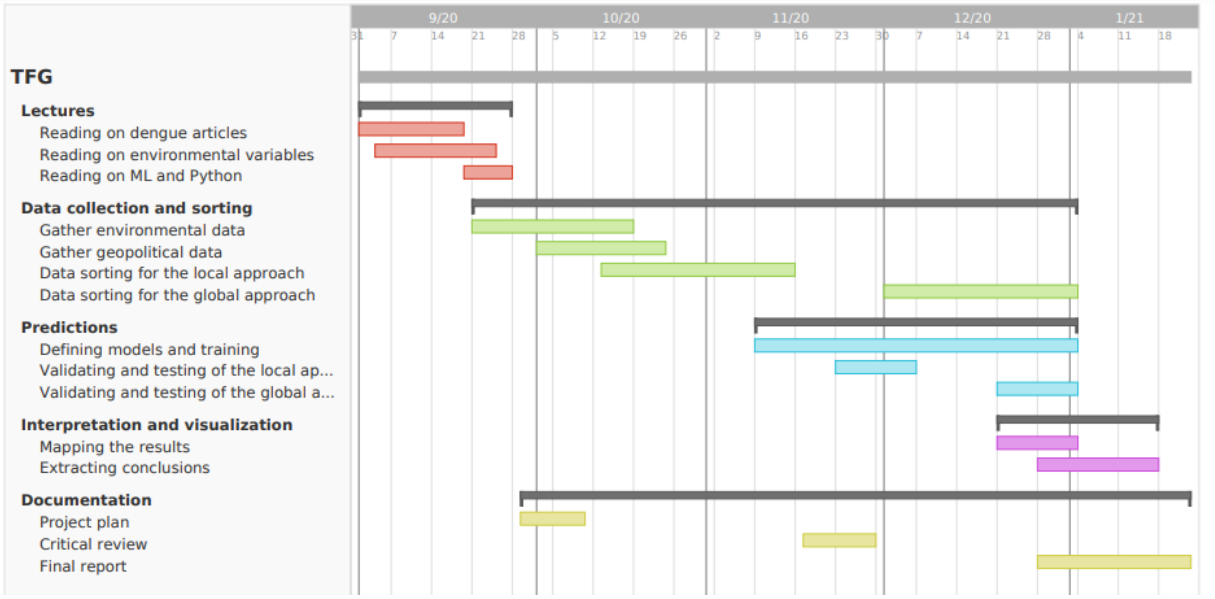

Figura 1. Diagrama de Gantt

1.4. Incidències i modificacions

La pandèmia de la Covid-19 no ha suposat un gran impediment pel desenvolupament del projecte, ja que la feina s'ha pogut fer còmodament a casa i anant algun dia puntual a la universitat.

Degut a la inexperiència amb la gran majoria dels conceptes (manipulació de dades de satèl·lits, georeferenciació i creació de bases de dades PostgreSQL entre altres) i les eines emprades hi ha hagut un retard significatiu en el desenvolupament de cada pas, cosa que ha limitat la profunditat de l'anàlisi global que es podria haver fet.

També, abans de crear els algoritmes de Machine Learning per tot el Brasil sencer, es va prendre la decisió de fer un anàlisi previ en una zona més reduïda per, d'aquesta manera,

garantir uns resultats. Així que es va establir una zona d'uns 78.400 km^2 al voltant de São Paulo per fer un primer model de predicció del risc epidemiològic.

El principal problema en la realització del projecte ha estat el format de les dades. Les dades socioeconòmiques i les dels casos de dengue estaven en format municipal. És a dir, cada municipi tenia el seu registre de número de casos mensual, població, índex de pobresa, educació, etc. Mentre que les dades climatològiques es presenten en un ràster en forma de malla. Llavors, el repte ha estat trobar un compromís entre els dos formats. En la primera aproximació s'ha intentat treballar amb una barreja entre els dos formats i en la segona s'han passat totes les dades al format municipal.

2. Variables i models per a la predicció del risc epidemiològic

El dengue és una malaltia produïda per un virus transmès pel mosquit *Aedes Aegypti*, i la seva presència es dóna sobretot als climes tropicals i subtropicals. Durant aquest segle el Brasil ha registrat més casos que qualsevol altre país del món. Els símptomes són relativament semblants als de la grip i pot afectar a persones de totes les edats. Els símptomes més comuns són: febre alta, nàusees i vòmits, erupcions a la pell, sensació de debilitat generalitzada, dolor muscular i articular, tos i mal de coll. En els casos més greus pot ser mortal, degut a la aparició de taquicàrdies, hemorràgies, alteració de la pressió arterial, insuficiència circulatòria o deshidratació.[1]

Aquest capítol es centra primer en descriure el virus i el seu origen, així com els aspectes dins de varis àmbits que poden tenir un efecte en la intensitat dels brots epidemiològics que genera. També es defineixen els algoritmes de Machine Learning que s'han utilitzat per crear un model predictiu de futurs brots al Brasil.

2.1. Característiques i propagació del dengue

Els mosquits són els principals transmissors de malalties infeccioses del món, i sobre totes les espècies de mosquit en destaca l'*Aedes Aegypti*. A part de ser el principal transmisor del dengue també transmet la chikungunya, el zika i la febre groga urbana.[2] No hi han vacunes per poder prevenir aquests virus, per tant una bona planificació és la millor solució, ja sigui, per exemple, fent una monitorització d'espècimens que es troben en una zona, o bé generant un model predictiu que pugui avisar de nous brots.

La transmissió del dengue s'ocasiona per un dels seus serotips immunològics. Aquests són el DEN-1, DEN-2, DEN-3, DEN-4 i DEN-5. Un cop rebuda la picada del mosquit s'inicia un període d'incubació que normalment dura dels 4 als 7 dies. Per a què els mosquits puguin transmetre la malaltia aquests l'han de rebre picant un humà ja infectat. En total han de passar unes tres setmanes perquè el virus es transmeti d'una persona infectada a una altra.

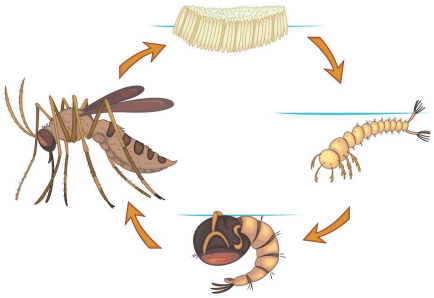

Figura 2. Fases de creixement de l'Aedes Aegypti.[3]

Hi han varies fases en la evolució d'un ou a un mosquit adult. Primer un mosquit femella diposita uns centenars d'ous, normalment en parets d'aigua estancada. En la segona fase els ous es converteixen en larva en un període de temps que va dels 5 als 11 dies. Després d'esdevenir una larva, el següent pas és evolucionar a una pupa en un procés que dura uns 3 dies. Finalment, en l'última fase la pupa s'obre i deixa sortir ja al mosquit adult

després de tres dies més. El mosquit adult no sol desplaçar-se més de 100 metres de la zona on ha nascut però en alguns casos pot arribar a desplaçar-se uns 3 km.

Per tant, l'*Aedes Aegypti* tarda aproximadament de 10 a 15 dies en convertir-se en adult. Aquest període de temps tan breu dificulta molt la contenció epidemiològica del dengue sota unes determinades condicions, com per exemple pluges abundants i breus que creïn numerosos estancaments d'aigua petits, que juntament amb temperatures elevades són ideals per la proliferació dels mosquits.

2.2. Paràmetres amb influència en la propagació del dengue

El desenvolupament de models fiables de predicció del risc epidemiològic requereixen d'una base de dades sòlida i àmplia. Avui en dia hi ha nombroses organitzacions i entitats públiques com les agències espacials americana (NASA) i la europea (ESA, amb el programa Copernicus), altres centres de distribució de dades com BEC (Barcelona Expert Center) i fins i tot particulars que comparteixen les seves bases de dades al públic de forma gratuïta i altruista. Moltes d'aquestes dades provenen de mesures des de satèl·lits d'observació de la Terra, que proveeixen les dades de manera periòdica i amb bona resolució. És gràcies a ells que projectes com aquest poden veure la llum sense una inversió de diners al darrere.

Les missions d'observació de la Terra, per tant, prenen mesures de forma periòdica de diferents paràmetres ambientals, climatològics i de l'estat del sòl i vegetació, així com, de l'estat dels mars i oceans. Les seves dades són molt importants per entendre els diferents procesos que ocorren sobre la Terra, així com per analitzar, monitoritzar l'evolució en diferents àrees del planeta i veure l'efecte del canvi climàtic. També ajuden a desenvolupar models climàtics i predictius cada vegada més precisos de catàstrofes naturals o degudes al efecte humà, com poden ser incendis, inundacions, sequeres, desglaç, etc.

En aquest projecte, es planteja l'ús de dades ambientals en obert de satèl·lit per al desenvolupament d'un model de risc de propagació del dengue a Brasil. En els següents apartats es detallen els que s'han utilitzat.

2.2.1. Mesura de variables ambientals amb teledetecció des de satèl·lit

Tal com s'ha mencionat, el mosquit *Aedes Aegypti* és el principal propagador del virus del dengue a Brasil. De manera que les condicions mediambientals que faciliten la seva proliferació i desenvolupament són determinants a nivell epidemiològic.

Els paràmetres climatològics que influeixen més en la proliferació de malalties virals transmeses per animals estan ben documentats i existeixen diferents estudis que en fan menció.En un estudi realitzat en la National Central University de Taiwan[4] l'any 2018, considera que alguns dels millors elements a tenir en compte com a indicadors per fer mapes de risc de malalties de transmissió vectorial són les precipitacions, l'índex d'ús del sòl, la temperatura superficial terrestre i l'índex normalitzat de diferència de vegetació (NDVI). Seguint aquesta línia, les variables triades per aquest projecte estan detallades a continuació.

NDVI

El *Normalized Difference Vegetation Index* (NDVI) és un índex que serveix per mesurar el creixement de la vegetació, el seu estat, determinar la superfície que ocupa i tenir controlada la producció de biomassa.

L'NDVI es calcula a partir de les reflectàncies en les bandes infrarroges properes (MIR) i vermelles (VIS) com indica la fòrmula 1.

$$NDVI = \frac{NIR - VIS}{NIR + VIS} \qquad (1)$$

Quan una planta és vigorosa reflexa molta radiació solar en l'infraroig proper i com a conseqüència projecta un NDVI elevat. Quan una planta no està en bon estat passa el contrari. De manera que l'NDVI queda sempre en l'interval entre -1 i 1. Definim que un NDVI inferior a zero implica una zona totalment artificial o d'aigua, entre 0 i 0.3 correspon a una zona de sòl amb vegetació marginal, i un valor de NDVI superior a 0.3 indica zones de vegetació. La figura 3 mostra el mapa mundial de NDVI pel mes de gener del 2010. S'ha obtingut a partir de les dades de MODIS de la NASA.

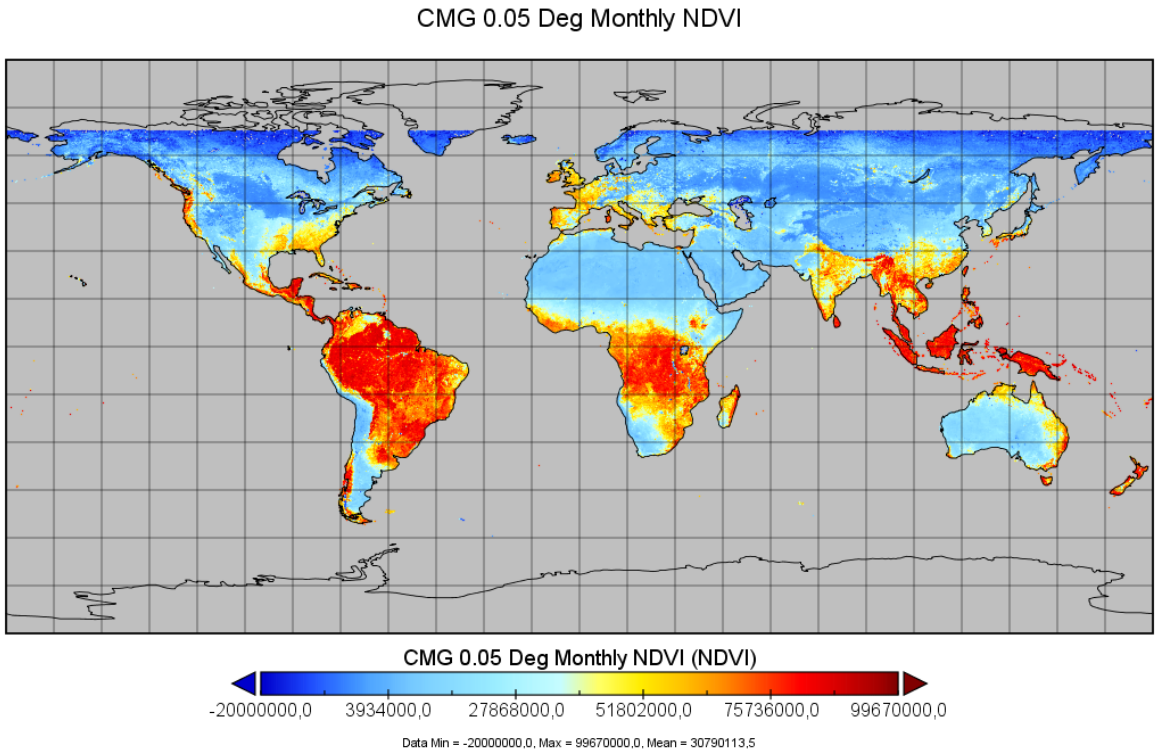

Figura 3. NDVI a nivell mundial al gener del 2010.

L'NDVI és una variable interessant a estudiar perquè en zones amb vegetació elevada prolifera la vida, ergo també les espècies de mosquits que potencialment poden transmetre malalties. La font utilitzada per extreure aquestes dades ha estat l'instrument de detecció MODIS,[5] que forma part d'un satèl·lit de la NASA.

Variable:	NDVI	**Resolució espacial**	**Resolució temporal**
Instrument:	MODIS	0.05 Deg	Mensual
Satèl·lit:	Terra	**Unitats**	**Format**
Organització:	NASA	-	hdf

Taula 1. Origen i format de les dades utilitzades NDVI.

La resolució espacial d'aquest instrument és de 0.05 graus, l'equivalent a uns 5.6 km en la latitud de la zona d'estudi. Mitjançant una interpolació s'ha pogut obtenir els valors promig per cada municipi de Brasil. Amb l'ajuda del QGIS s'han pogut visualitzar. Quan s'analitza l'evolució del NDVI d'un any es detecten canvis estacionals notables, sobretot en àrees de cultius.

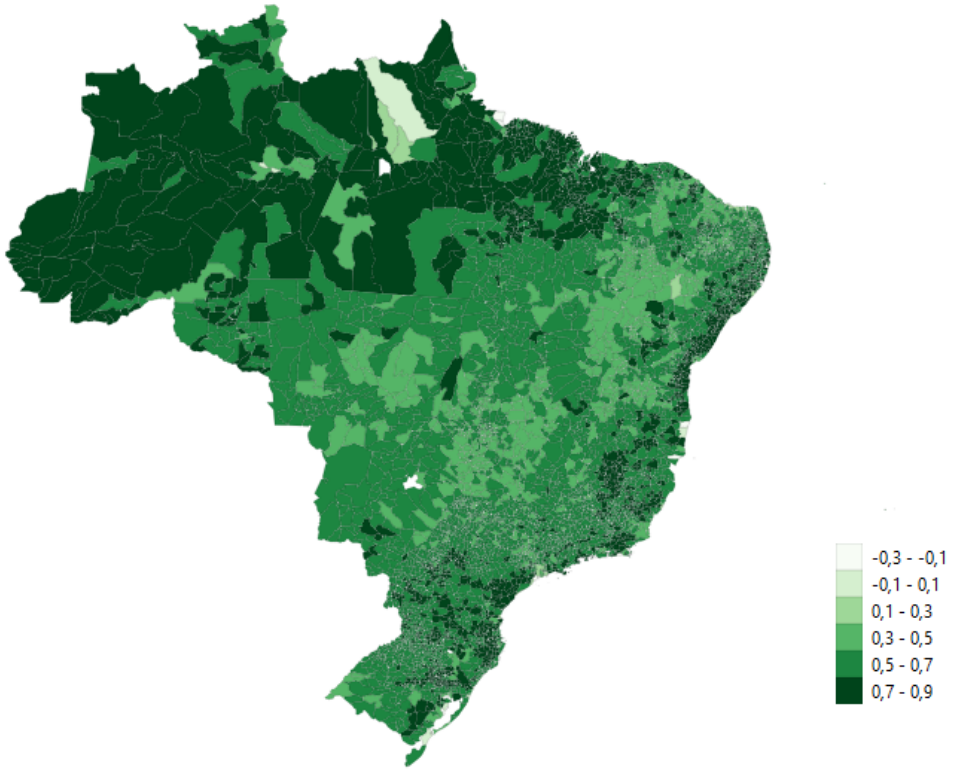

Figura 4: NDVI a nivell municipal al juny de 2010 a partir de les dades de MODIS.

NDWI

A la literatura es presenten dues definicions del *Normalized Difference Water Index* (NDWI). Una s'utilitza per mesurar els canvis del contingut d'aigua de les fulles, segons l'estudi de Bo-cai Gao a l'any 1996[6] i l'altre que mesura l'estat de les masses d'aigua, definit per S. K. McFeeters l'any 1996[7]. És una variable molt útil per controlar el risc de sequeres, incendis i inundacions.

El primer es calcula a partir de la reflectància de diferents bandes de l'espectre electromagnètic. Aquestes bandes són l'infraroig proper (NIR) i l'infraroig mitjà tal com indica la fórmula 2.

$$NDWI = \frac{NIR - MIR}{NIR + MIR} \qquad (3)$$

La seva obtenció no és directa a partir de la font de dades utilitzada sinó que s'extreu a partir dels dos següents mapes.

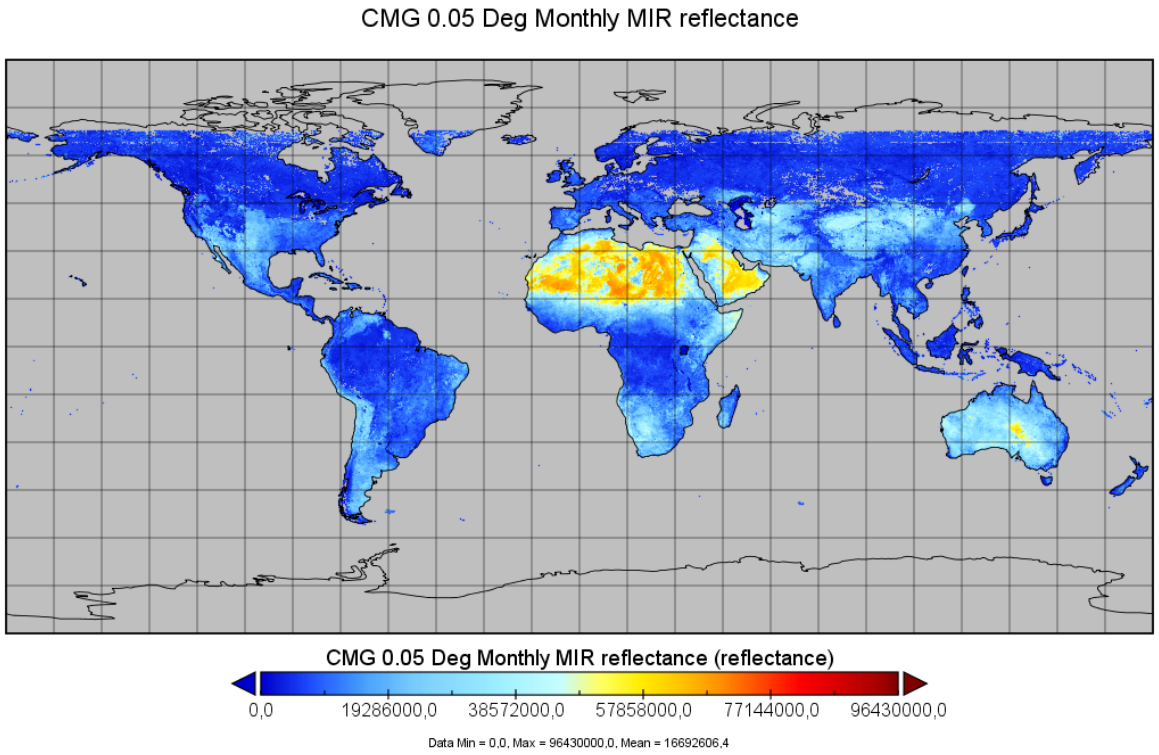

FIgura 5. Reflectància MIR mundial al gener de 2010.

La NDWI varia entre el -1 i 1, depenent del tipus de vegetació, cobertura, etc. Els valors alts de NDWI corresponen a un alt contingut d'aigua en la planta i al seu recobriment. Com que sabem que l'*Aedes Aegypti* posa els ous a estancacions d'aigua la NDWI és una variable clau per anticipar-se a poblacions elevades del mosquit.

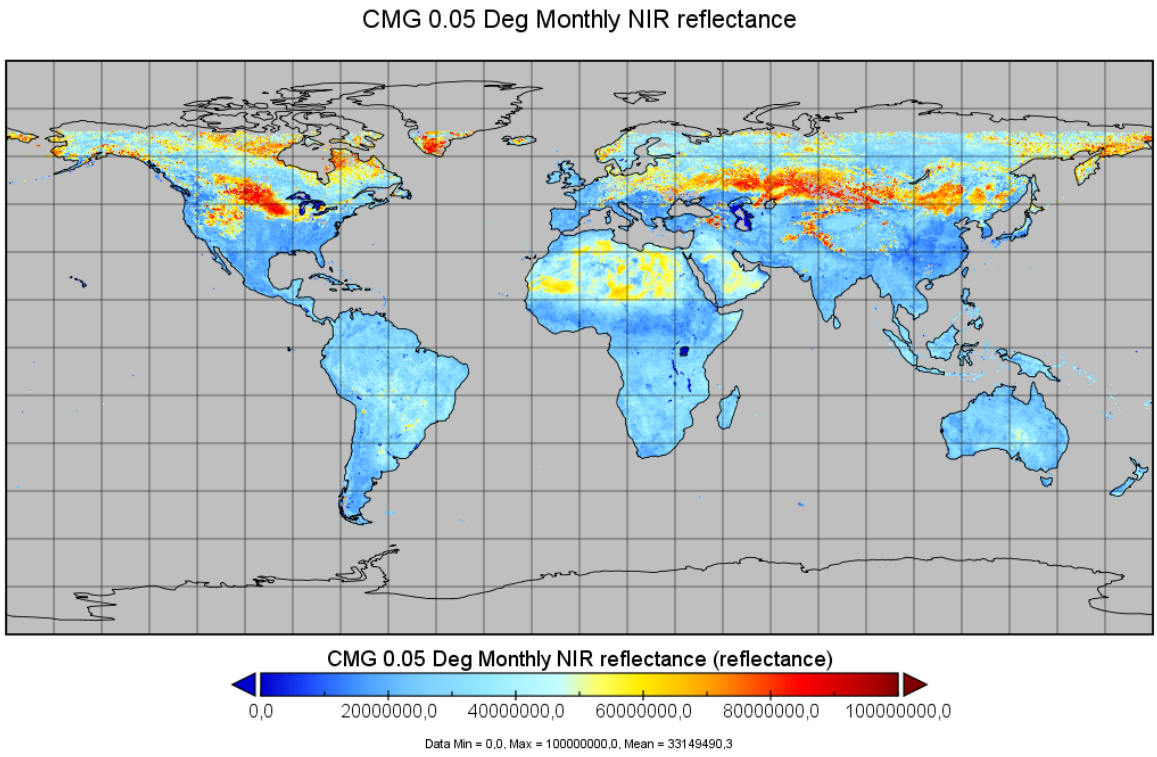

Figura 6. Reflectància NIR al gener de 2010.

Tal com la NDVI, la NDWI s'ha obtingut a partir del satèl·lit Terra de la NASA. Les especificacions de les dades queden detallada a continuació. Aquestes es distribueixen amb la plataforma LP DAAC[8] de la NASA.

Variable:	NDWI	**Resolució espacial**	**Resolució temporal**
Instrument:	MODIS	0.05 Deg	Mensual
Satèl·lit:	Terra	**Unitats**	**Format**
Organització:	NASA	-	hdf

Taula 2. Origen i format de les dades utilitzades NDWI.

Humitat del sòl

La humitat del sòl (SM, *Soil Moisture* en anglès) és la quantitat d'aigua que queda entre partícules de sòl. Actualment existeixen dos satèl·lits que treballen a la freqüència de 1.4 GHz dedicats a mesurar la humitat tel sòl: SMOS (*Soil Moisture and Ocean Salinity*) de l'ESA, i SMAP (*Soil Moisture Active and Passive*) de la NASA. A 1.4 GHz els radiòmetres són capaços de mesurar aquesta variable fins a uns 10 cm de profunditat.[9] Degut al procés reproductiu de l'*Aedes Aegypti* aquesta mesura podria ser important, sobretot en

àrees de boscos i agrícoles per a mesurar-ne el risc d'un futur brot de dengue. La figura 7 mostra la humitat del sòl a 25 km de resolució extreta amb les dades proporcionades pel BEC del juny del 2010.

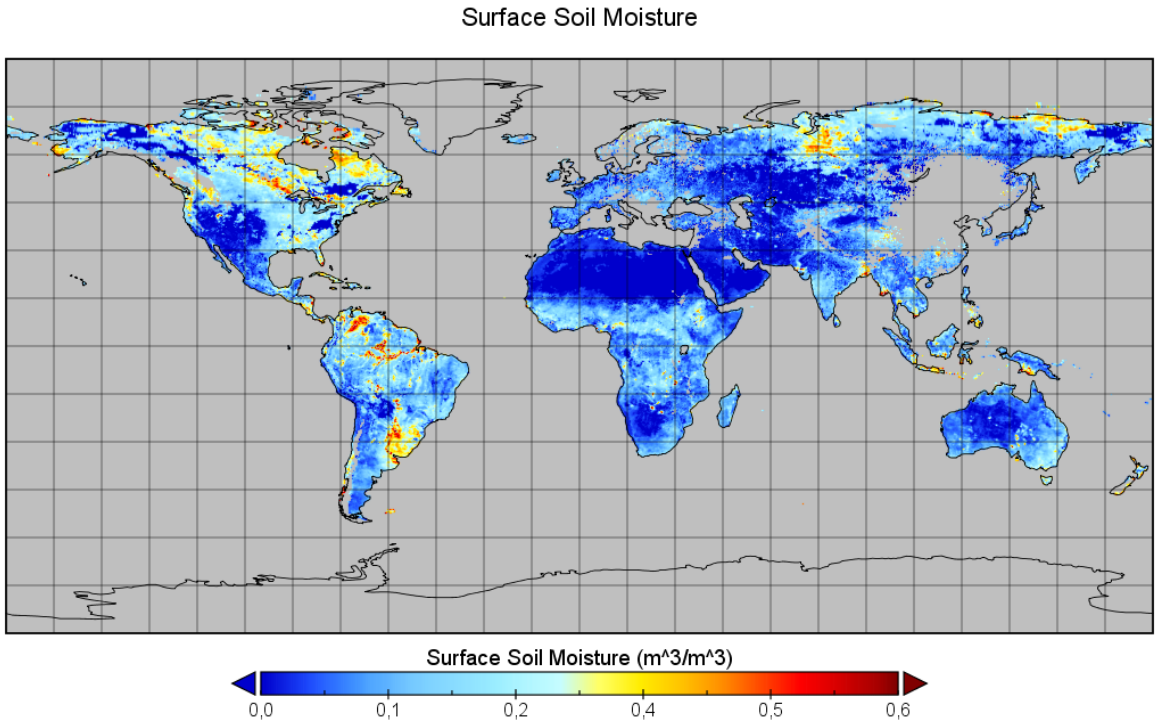

Figura 7. SM al juny de 2010.

Les dades utilitzades d'aquesta variable s'han obtingut del centre de distribució de dades Barcelona Expert Center (BEC) tal com es detalla a la taula 3. Aquest centre de producció i distribució de dades és fruit d'una iniciativa conjunta entre el grup de radiometria de la UPC i l'Institut de Ciències del Mar (ICM).

Variable:	SM	**Resolució espacial**	**Resolució temporal**
Instrument:	MIRAS	25 km	Mensual
Satèl·lit:	SMOS	**Unitats**	**Format**
Organització:	BEC	m^3/m^3	netCDF

Taula 3. Origen i format de les dades utilitzades SM.

Temperatura superficial terrestre

La temperatura superficial del sòl, en anglès *Land Surface Temperature* (LST), igual que el NDVI i NDWI, en aquest projecte s'ha obtingut del sensor MODIS de la NASA. Es mesura en Kelvins i és important ja que la temperatura ideal per la proliferació del mosquit és entre 26ºC i 28ºC.[10] En canvi a una temperatura inferior a 10ºC la capacitat reproductiva és pràcticament nula. Les temperatures més altes de 31ºC acceleren

l'envelliment del mosquit, reduint així el període de temps que pot resultar infecciós. La figura 8 mostra un mapa global de la temperatura mitjana mensual pel gener del 2010. Com que coincideix amb l'hivern de l'hemisferi nord les temperatures són baixes (colors blaus) i altes a l'hemisferi sud (colors vermells). A més la taula 4 conté un resum de la informació sobre la missió i paràmetres utilitzats en el projecte per aquesta variable.

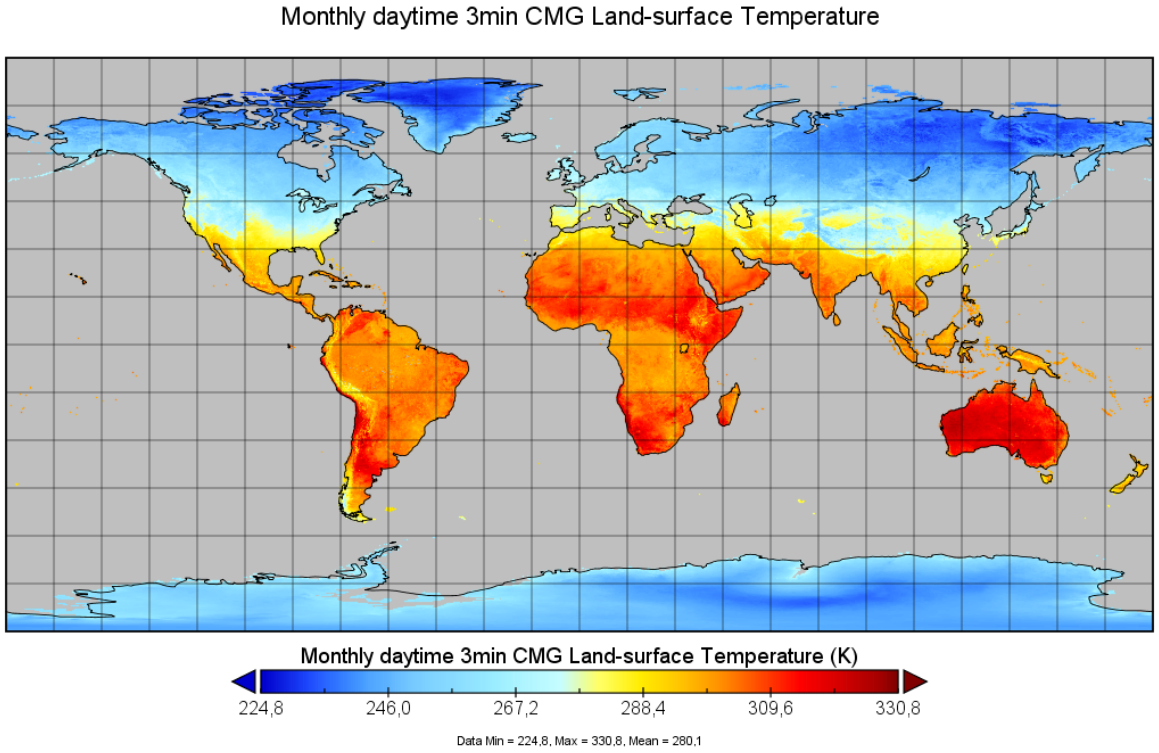

Figura 8. Temperatura superficial terrestre diurna mitjana del mes de gener de 2010.

Variable:	Temperatura	**Resolució espacial**	**Resolució temporal**
Instrument:	MODIS	0.05 Deg	Mensual
Satèl·lit:	Terra	**Unitats**	**Format**
Organització:	NASA	K	hdf

Taula 4. Origen i format de les dades utilitzades de la temperatura superficial.

Precipitacions acumulades

La informació de les precipitacions acumulades en una zona també pot ser una de les variables que ens pot ajudar a determinar el risc de propagació de la malaltia, doncs pot influir en el cicle reproductiu i de creixement de l'*Aedes Aegypti*. Dos satèl·lits dedicats a produir aquesta variable són TRMM (*Tropical Rainfall Measuring Mission)* i GPM (*Global Precipitation Measurement)*, gestionats per la NASA i la JAXA (Agència Japonesa d'Exploració Aeroespacial). El TRMM porta inoperatiu des del 2014 mentre que el GPM és el satèl·lit que ha seguit funcionant fins a dia d'avui. A la pàgina web

https://gpm.nasa.gov/missions/GPM se'n troba informació d'aquestes missions i a la taula 5 es resumeix les característiques de les dades utilitzades en aquest projecte. La figura 9 mostra un mapa d'exemple de les precipitacions acumulades (en mm) pel mes d'abril al Brasil. La representació s'ha fet a nivell municipal i els colors més foscos de blau indiquen quantitats més elevades de pluja.

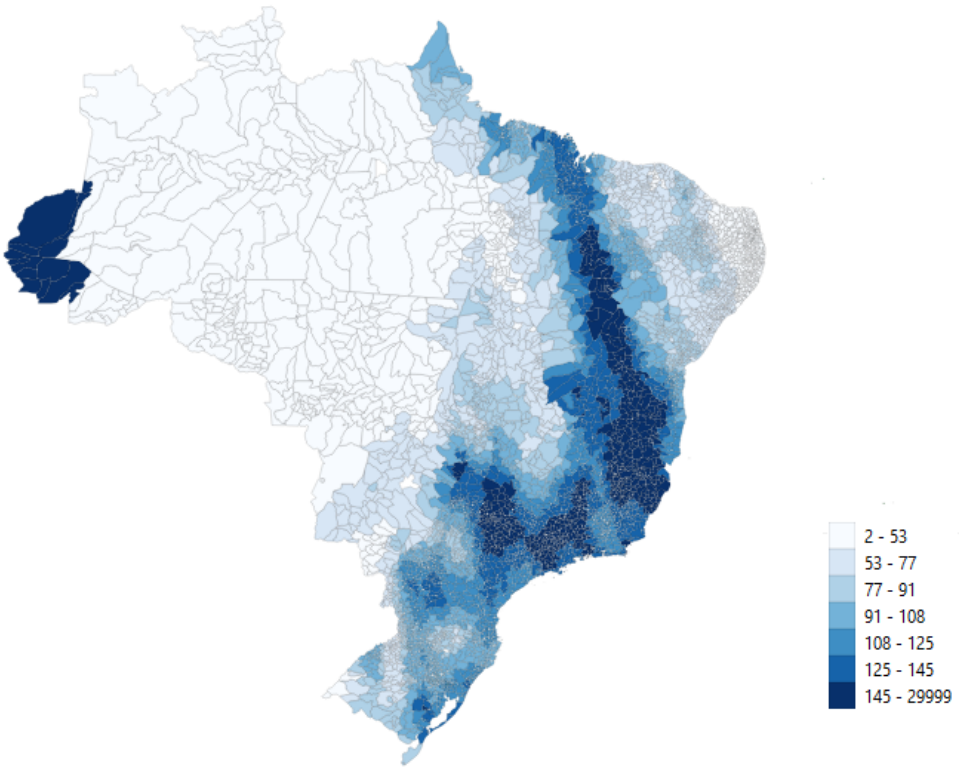

Figura 9. Precipitacions acumulades (mm) a nivell municipal al febrer de 2010.

Variable:	Precipitacions	**Resolució espacial**	**Resolució temporal**
Instrument:	Múltiples	0.1 Deg	Mensual
Satèl·lit:	GPM	**Unitats**	**Format**
Organització:	NASA i JAXA	mm	GeoTIFF

Taula 5. Origen i format de les dades utilitzades de les precipitacions acumulades.

2.2.2. Variables socioeconòmiques

Les variables ambientals són útils però no suficients a l'hora de realitzar la predicció del risc epidemiològic. L'ésser humà, intencionadament o no, té molta influència en la proliferació de vida tant en el medi natural com urbà. Per exemple, en un entorn de pagès no sería extrany la presència de petits estancaments d'aigua, ja sigui com a abeuradors del bestiar o basses de regadiu pel conreu i que en els barris desafavorits de les ciutats, com les faveles del Brasil on no hi ha distribució d'aigua potable també és habitual trobar petits contenidors per emmagatzemar-la. Així doncs, aquests mateixos estancaments d'aigua són ideals per a la reproducció de l'*Aedes Aegypti*.

Per aquest motiu en el projecte s'ha buscat una base de dades d'accés lliure que proporcioni uns indicadors a nivell municipal relatius al nivell econòmic i cultural de les persones que pugui ser d'ajuda per predir brots epidemiològics del dengue.

S'ha utilitzat dues bases de dades,[11] una ha estat creada per Cristina Parada i consisteix en la recopilació de dades proporcionades pel govern de Brasil, entre altres institucions. A continuació es detallen les variables que se n'han extret. Per altra banda, el nombre de casos de dengue enregistrats per mesos durant el període d'estudi s'ha obtingut d'una base de dades proporcionada per Hellen Gurgel, de la Universitat de Brasilia, amb qui s'està col·laborant per aquesta línia de recerca i que les obté directament del Ministeri de Sanitat del Brasil. A continuació es descriuen les variables d'aquest tipus utilitzades en l'estudi.

Risc epidemiològic

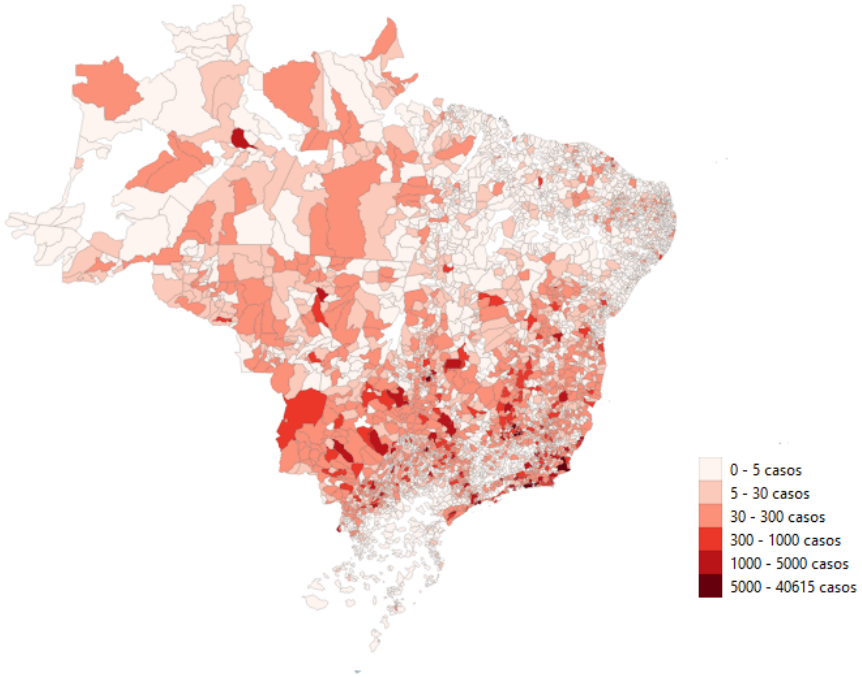

Figura 10. Nombre de casos de dengue confirmats a nivell municipal al febrer de 2010.

La figura 10 mostra el nombre de casos de dengue confirmats al febrer de 2010. L'índex de risc epidemiològic del dengue s'ha obtingut a partir dels registres de casos a nivell municipal proporcionat per Hellen Gurgel, de la Universitat de Brasilia, i la base de dades de Cristina Parada, que en aquest cas indica la població de cada municipi. L'índex de risc epidemiològic de cada municipi s'ha definit aplicant les següents fórmules.

$$Casos_{100.000Hab} = N_{casos} \frac{100.000\ Hab}{Població} \qquad (3)$$

$$Risc = min\left\{\frac{Casos_{100.000Hab}}{500},\ 1\right\} \qquad (4)$$

S'estableixen dues escales de risc detallades a continuació. Aquestes escales estan basades en l'estudi "*Dengue outlook for the World Cup in Brazil"* publicat per *The Lancet*.[12] Així doncs aquests índexs serveixen tant per definir el risc epidemiològic en el mateix moment que es prenen les altres mesures, i s'introdueix als models com una dada més, i també per definir el risc a predir del mes següent.

Índex A	**Rang de valors**	**Índex B**	**Rang de valors**
Mínim	*< 100 casos$_{100khab}$*	Baix	*< 200 casos$_{100khab}$*
Baix	*Entre 100 i 200 casos$_{100khab}$*	Mitjà	*Entre 200 i 300 casos$_{100khab}$*
Mitjà	*Entre 200 i 300 casos$_{100khab}$*	Alt	*> 300 casos$_{100khab}$*
Alt	*Entre 300 i 400 casos$_{100khab}$*		
Molt alt	*> 400 casos$_{100khab}$*		

Taula 6. Índexs de risc aplicats en el desenvolupament dels models.

Índex de Desenvolupament Humà Municipal (IDHM)

L'IDHM és la variant a nivell municipal de l'índex de desenvolupament humà (IDH). L'IDH és una estadística d'ús estès amb l'objectiu de proporcionar una mesura que representi el nivell o qualitat de vida d'un lloc. Va ser creat l'any 1990 per Mahbub Ul Haq i Amartya Sen pel Programa de les Nacions Unides per al Desenvolupament de la ONU.[13]

Aquest índex es calcula a partir de la mitjana de tres altres índex. Primer, la durada i la qualitat de vida, obtingut a partir de la esperança de vida del lloc. El segon és el nivell d'instrucció i educació. Finalment, el tercer és el nivell de vida a partir del PIB per càpita. Els càlculs utilitzats per obtenir cadascun d'aquests índexs es mostren a les fórmules 5, 6 i 7, i la figura 11 representa la mitjana d'aquests en cada municipi.

$$Index\ de\ l'esperança\ de\ vida = \frac{EV - 20}{85 - 25} \quad (5)$$

$$Index\ d'educació = \frac{2}{3} IAA + \frac{1}{3} TA \quad (6)$$

$$Index\ PIB = \frac{log(PIB_{PC}) - log(100)}{log(40.000) - log(100)} \quad (7)$$

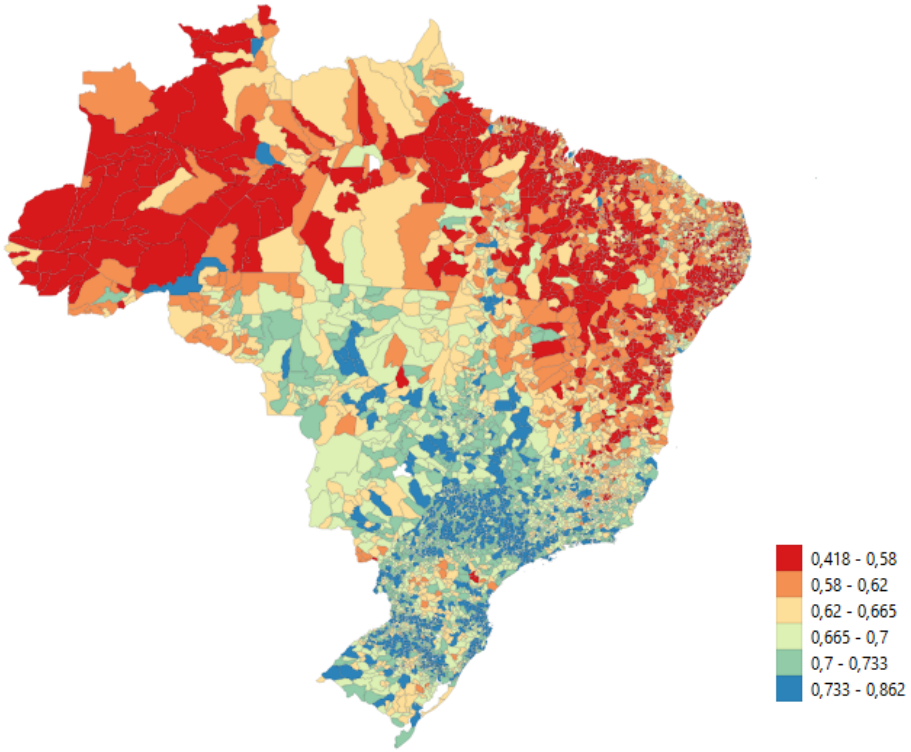

Figura 11. IDHM de cada municipi del Brasil.

Tipus de sòl

En aquest estudi s'ha considerat important incorporar com a paràmetre el tipus de municipi que s'està estudiant. La propagació de l'*Aedes Aegypti* en un entorn rural o en un entorn urbà ha de ser previsiblament diferent. És per això que s'ha incorporat una variable que ens descriu la tipologia de cada un dels municipis de Brasil. S'ha fet la classificació en cinc categories en l'espectre de nivell d'urbanització:

- Rural remot.
- Rural adjacent
- Intermediari remot
- Intermediari adjacent
- Urbà

Cal notar que encara que un municipi estigui indicat com a terreny urbà no vol dir que tot el terme municipal estigui urbanitzat, sinó que el nucli, on viu la majoria de la població, sí que ho està. La figura 12 mostra aquesta classificació.

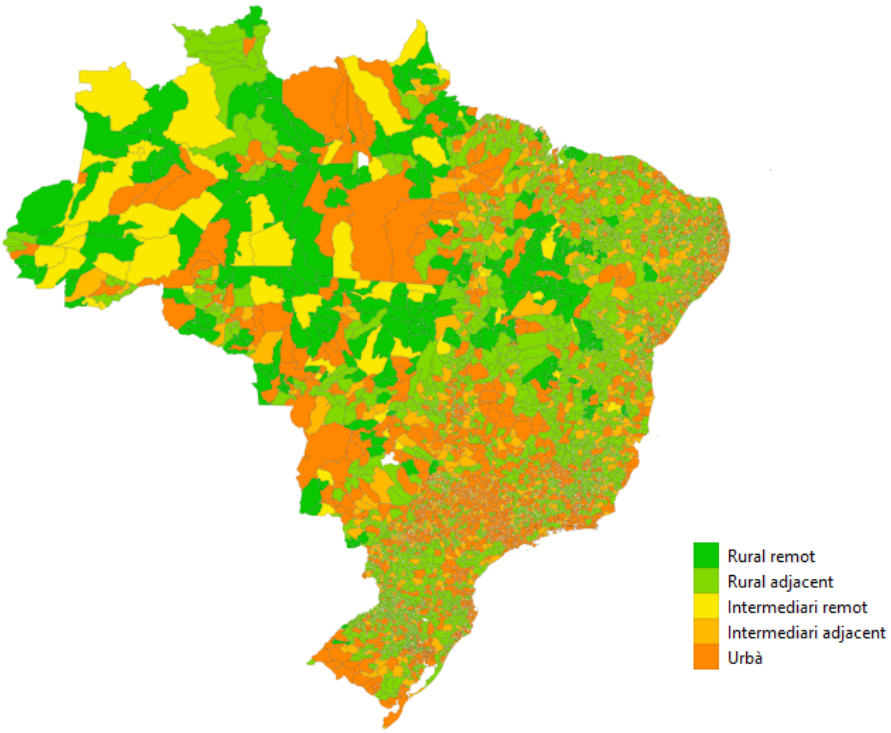

Figura 12. Tipologia urbana de cadascun dels municipis de Brasil.

Altres variables no incloses

En la mateixa base de dades de la que s'han extret les variables socioeconòmiques esmentades anteriorment se'n troben unes de diferents que podrien resultar interessants incorporar en un futur model predictiu del risc epidemiològic més ampli. Algunes d'elles són:

- Hectàrees de conreu.
- Número d'entitats dedicades a la sanitat i serveis socials.
- Població desglossada per edats.
- Producte interior brut per càpita.

2.3. Models predictius

Un cop recopilades les dades ambientals i socioeconòmiques de tot el país es pot veure a simple vista que hi ha una correlació amb aquestes dades i el número de casos de dengue a nivell local, però no és senzill crear un model matemàtic per relacionar-ho directament. És per això que aplicar processos d'aprenentatge autònom és una bona opció per poder generar prediccions de manera fiable.

Aquest aprenentatge es fa a través del desenvolupament d'algorismes i metodologies que generen una evolució en les màquines que els realitzen, i un exemple n'és el *Machine Learning*.

Els algoritmes de *Machine Learning* es poden dividir en les següents categories:

- Aprenentatge supervisat: es crea una relació entre unes dades d'entrada i unes de sortida.
- Aprenentatge no supervisat: es disposa d'exemples d'entrada, però no hi ha unes dades de sortida per crear una relació directa, sinó que l'algorisme mateix ha de trobar patrons per poder generar etiquetes.
- Aprenentatge semisupervisat: es combinen les dues metodologies anteriors amb dades marcades i no marcades.
- Aprenentatge per reforç: després d'una decisió presa per l'algoritme se l'hi aplica una recompensa o una penalització segons els resultats.

Com és lògic, degut al tipus de disposició de dades d'aquest projecte, es crearan algorismes d'aprenentatge supervisat, ja que tant les dades d'entrada com les de sortida han estat recollides prèviament.

Una altra manera de dividir els algoritmes de *Machine Learning* és definir si són de classificació o regressió. La variable de sortida dels algorismes de classificació és una classe, és a dir, una categoria arbitrària que millor defineix un concepte. En canvi la variable a predir en els algorismes de regressió és un valor numèric, per tant té infinites possibilitats i les unitats de mesura de la precisió de la predicció seran diferents que en el cas anterior.

Les dades recollides per desenvolupar el projecte presenten una gran variança, i en conseqüència s'ha decidit transformar els valors de les variables de risc en diferents classes (segons es detalla en l'apartat 2.2.2.), i enfocar la predicció com una classificació o bé una regressió (amb posterior classificació) segons es cregui convenient. La variança és elevada degut a que s'intenta generar un model que englobi els 5.564 municipis del Brasil amb les dades de varis anys consecutius.

A continuació es detallen els algorismes utilitzats.

2.3.1. Decision Trees

Els arbres de decisió (*decision trees* en anglès) són un dels models predictius més coneguts en estadística, *data mining*, i *Machine Learning*. Els *decision trees* són reconeixibles per la seva forma de copa d'arbre inversa que va dividint-se en branques. Es col·loca el paràmetre més important a la part superior i mitjançant diverses

comparacions basades en mesures estadístiques, es creen subespais dins del bloc de dades segons compleixin o no les condicions que s'estableixen.

Els *decision trees* es poden dividir en dues categories: *classification trees* i *regression trees*. Pel desenvolupament d'aquest projecte s'utilitzaran els *classification trees*.

Segons s'estableixen els umbrals de decisió en l'esquema destaquen dos algoritmes dins dels *decision trees*. Aquests són l'algorisme C4.5 i el CART.

- **Algorisme C4.5**: està basat en l'algorisme previ, ID3, i el concepte d'entropia. Les dades d'entrenament es divideixen en mostres $S = s_1, s_2, s_3, \ldots$ ja classificades. Cada mostra s_i consisteix en un vector on cada coordenada representa el valor d'un dels camps del bloc de dades.

  A cada node de l'arbre, el C4.5 decideix quin atribut és el que genera millors divisions dins de les dades segons el criteri del guany d'informació.

$$IG(Dp,\ f) = I(Dp) - \frac{N_{esquerra}}{N} I\,(D_{esquerra}) - \frac{N_{dreata}}{N} I\,(D_{dreta}) \qquad (8)$$

  On *Dp* correspon a les dades totals a classificar. *I* correspon al criteri de decisió (Gini o entropia) i *N* és el nombre total de mostres.

- **Algorisme CART** (*Classification And Regression Trees*): segueix una metodolgia semblant al C4.5. Es creen normes basades en els valors de les variables de manera que es seleccioni la millor divisió. Un cop una norma es crea i un node es divideix en dos, el mateix procés s'aplica progressivament als nodes inferiors. Les divisions s'aturen quan l'algorisme detecta que no hi ha una millora possible. Els dos criteris més utilitzats per generar aquests subespais de forma precisa són l'índex Gini (9) i l'error quadràtic mitjà (10).

$$Gini = 1 - \sum_{i=1}^{N} p_i^{\,2} \qquad (9)$$

$$mse = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2 \qquad (10)$$

2.3.2. Random Forests

Els *decision trees* són una solució obvia i simple pels casos on sigui necessari fer una predicció, i més quan aquests no generen cap error amb les dades utilitzades, però cal recordar que l'objectiu del *Machine Learning* és generalitzar al màxim les prediccions, fins i tot amb dades que encara no es tenen. És per això que els *random forests* destaquen sobre els *decision trees*, al ser-ne una versió més elaborada.

L'*overfitting* és un problema comú dels *decision trees* ja que al ser un model molt flexible tendeixen a "memoritzar" les dades. Dit d'altra manera, un model flexible tendeix a ser un model amb una alta variància. En canvi un model inflexible tendirà a tenir un alt biaix degut a que el seu anàlisi de les dades estarà molt definit.

Una possible solució a aquest problema dels *decision trees* és crear-ne varis i agrupar-los, en el que és conegut com a *random forest*, que es caracteritza per:

- Agrupació aleatòria de mostres de dades quan es crea cada arbre.
- Selecció aleatòria de variables quan es divideixen els nodes.

La predicció global s'aproxima a partir de la predicció de cada un dels arbres, que com que cadascun d'ells ha estat entrenat amb una mostra de dades diferent, la variància serà menor.

L'altre concepte principal que defineix els *random forests* és que en cada arbre només un subconjunt de variables és considerat. Generalment aquest número de tipus de mostres de cada arbre s'obté utilitzant l'equació 11, essent *S* el número total de tipus de mostres.

$$S_i = \sqrt{S} \qquad (11)$$

2.3.3. Neural Networks

Les xarxes neuronals o *neural networks* són una eina dins del que es coneix com a *deep learning*, o aprenentatge profund, que es basa en la agrupació per capes d'elements d'activació. Els més bàsics són coneguts com a *perceptrons*.

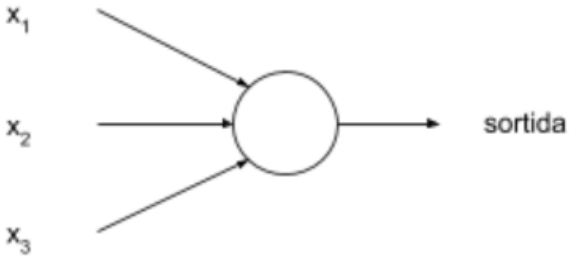

Figura 13. Perceptron simple.

Cada *perceptron* té un nombre indeterminat d'entrades, x_1, x_2,... I a cada entrada se li assigna el valor corresponent a un pes que servirà per calcular el valor de la sortida, w_1, w_2,... on a partir de les dades d'entrenament s'anirà ajustant aquests pesos progressivament de manera que generin una predicció més acurada. També s'ha de tenir en compte el biaix b_1, b_2,... de cada variable ja que es parteix de la suposició que no totes tenen la mateixa importància. El valor de sortida es calcula segons:

$$sortida = \begin{cases} 0 & si\ w \cdot x + b \leq llindar \\ 1 & si\ w \cdot x + b > llindar \end{cases} \qquad (12)$$

També cal notar que aquesta no és la única funció d'activació, sinó que n'hi ha de molts tipus, alguns exemples queden detallats en la següent taula. La complicació de les xarxes neuronals ve de saber triar quina és la funció d'activació més adequada per cada situació i també les dimensions de la xarxa (nombre de capes ocultes, nombre de neurones per cada capa, etc).

Sigmoid	**ReLU**	**Leaky ReLU**
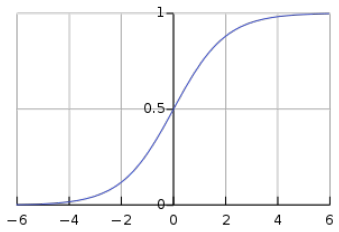	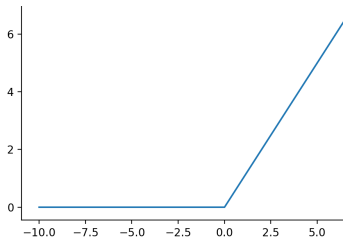	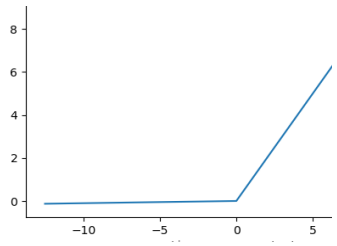
Figura 14. Sigmoide.	*Figura 14. Rectified Linear Unit.*	*Figura 16. Leaky ReLU.*
$\sigma(z) = \frac{1}{1+e^{-z}}$	$\sigma(z) = \begin{cases} 0 & si\ z < 0 \\ z & altrament \end{cases}$	$\sigma(z) = \begin{cases} 0.01z & si\ z < 0 \\ z & altrament \end{cases}$

Taula 7. Algunes funcions d'activació de les xarxes neuronals.

Tal com s'ha comentat anteriorment el procés d'aprenentatge vé donat pel constant ajustament dels valor de pes i biaix, i això s'aconsegueix minimitzant els errors. La retropropagació[14] (*backpropagation* en anglès) és un algorisme que s'utilitza per aquest propòsit. La retropropagació divideix l'error recursivament a través de les múltiples connexions de la xarxa calculant el gradient de la funció de cost associada a un estat.

Una altra mètrica utilitzada és el *learning rate*, que correspon a la mida dels passos que el model realitza quan intenta ajustar un error en cada observació. Un *learning rate* elevat redueix el temps d'entrenament, però pot comportar problemes de precisió.

2.4. Format i processament de les dades

El procés de recollida de les dades i el seu tractament ha estat llarg degut a les diferents procedències (NASA, Universitat de Brasilia, repositoris públics, etc) i les discrepàncies de formats (diferents precisions, tipus d'arxius, sistemes de coordenades, etc). Aquest procés es pot resumir amb la figura 14.

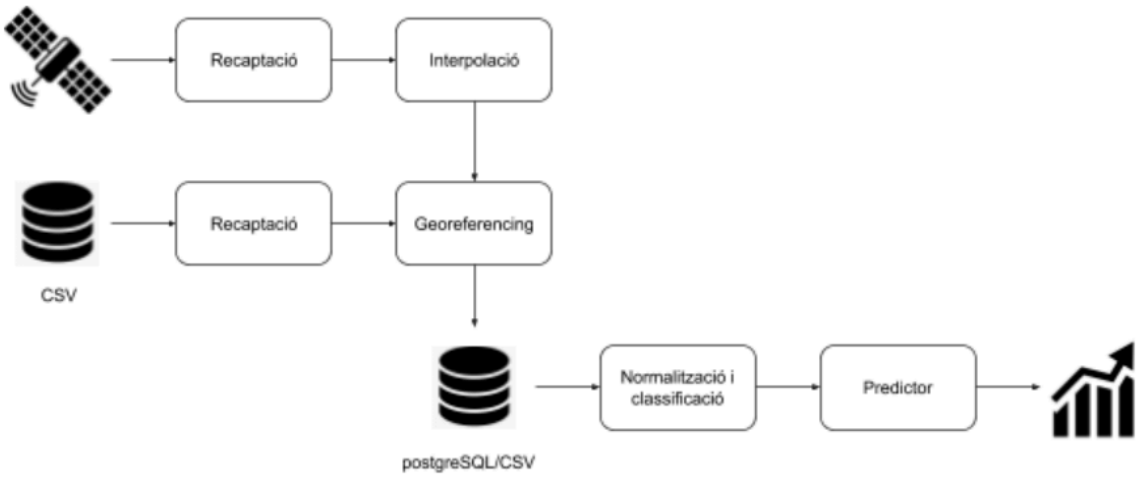

Figura 14. Procés aplicat a les dades utilitzades.

El primer pas ha estat recollir les dades de satèl·lits pertinents i realitzar un seguit de interpolacions i transformacions per poder-les utilitzar indistintament i relacionar-les amb les dades municipals. Llavors s'ha georeferenciat inversament aquestes dades amb un algorisme simple que s'ha creat, de manera que a cada municipi li correspon un número identificatiu únic.

Totes les dades s'han ajuntat i desat en una base de dades postgreSQL, que inclou una extensió postGIS amb les formes digitalitzades de cada municipi per tal de facilitar-ne la visualització amb QGIS.

3. Desenvolupament dels models predictius de risc epidemiològic

Aquest apartat conté l'explicació de cadascun dels processos que s'han dut a terme per a la creació dels diversos models predictius a partir dels conceptes detallats en l'apartat anterior, des de la recaptació de les dades fins a la visualització dels resultats.

3.1. Tractament de les dades

Les dades s'han organitzat tal que cada entrada representa un espai, ja sigui un punt d'un ràster en forma de malla o bé un terme municipal. Dins de cada entrada, que consisteix en una fila dins de la base de dades, es troba tota la informació tant ambiental com socioeconòmica d'aquest espai.

El primer que s'ha fet ha estat crear una base de dades on la discriminació temporal s'expressa en diferents columnes. Aquest format s'ha anomenat com a *raw*. Les dimensions són variables. Per exemple, si s'agafen les dades de quatre anys la base de dades tindrà 439 columnes.

ID municipal	latitud	longitud	població	àrea	IDHM	Tipologia urbana	N casos gener 2010	...	N casos desembre 2013	risc gener 2010	...	risc desembre 2013	NDVI gener 2010	...	NDVI desembre 2013	NDWI gener 2010	...	NDWI desembre 2013

Taula 8. Distribució de les fileres a la base de dades raw.

El format *raw* és ideal per manipular les dades de forma ordenada degut a la seva disposició, i també per poder-les visualitzar en un mapa amb el QGIS. En aquest cas cada municipi o punt del ràster correspon a una fila única que conté tots els valors que el representen. El problema sorgeix quan es vol introduir aquestes files a un model predictiu, ja que caldria realitzar moltes transformacions per tal de que les dades d'entrada siguin llegibles i útils per aquests. És per això que s'ha creat una altra base de dades a partir de la anterior, anomenada *cooked*.

ID municipal	latitud	longitud	mes	NDVI	NDWI	temperatura diurna	temperatura nocturna	Humitat	Precipitacions	IDHM	Tipologia urbana	risc actual	risc a predir

Taula 9. Distribució de les fileres a la base de dades cooked.

Llavors, la introducció de les dades en aquest format en els diferents models predictius és directa, només descartant les columnes que no interessen, ja que els valors corresponents a les variables ambientals i socioeconòmiques ja han estat seleccionats per una localització i mes concret, i també normalitzats. Degut a que cada localització, en aquest format, té tantes files com mesos abarca l'anàlisi, la base de dades pot arribar a tenir centenars de milers d'entrades o files.

3.1.1. Compatibilització de formats

Un dels principals problemes i reptes encarats durant la realització del projecte ha estat les discrepàncies entre els formats de les dades ambientals i socioeconòmiques. Algunes d'aquestes dades estan referenciades amb el número d'identificació que tenen els 5.564 municipis del Brasil, mentre que altres, les variables ambientals, prenen com a referència un sistema de coordenades geogràfiques, que utilitza les coordenades angulars longitud i latitud terrestre.

Un altre inconvenient ha estat el format que fa servir la NASA per distribuir les dades de satèl·lit de les variables NDVI, NDWI i temperatura. Aquest format és HDF4, una versió més antiga de l'actual HDF5. Degut a la seva antiguitat no s'ha trobat cap paquet de Python viable per poder-hi treballar. Per aquest motiu per extreure les dades d'aquests mapes s'ha fet servir MATLAB, que sí que suporta HDF4.

3.1.2. Georeferenciació inversa

La georeferenciació[15] és un conjunt de tècniques amb l'objectiu d'identificar objectes geogràfics. Com a objecte geogràfic s'entén qualsevol element o estructura que es pugui relacionar d'alguna manera amb una localització geogràfica, com per exemple un punt d'interès (carreteres, places, ponts, etc). Per altra banda una localització geogràfica és una entitat que representa un espai que es pot definir en diverses dimensions: cap dimensió (punts), una dimensió (línies), dues dimensions (àrees) i tres dimensions (volums).

Llavors, s'entén com a georeferenciació inversa el procés contrari. És a dir, a partir d'una localització geogràfica obtenir-ne la informació de l'objecte geogràfic que l'hi correspon. En aquest projecte ha interessat determinar per a cada punt (longitud i latitud) quin és el número identificatiu del municipi que el conté.

Per aquest propòsit s'ha creat un algorisme que a l'introduir un seguit de coordenades a dins d'un DataFrame de Pandas (una biblioteca de manipulació i anàlisi de dades de Python), aquest retorna el número d'identificació municipal que els hi correspon. Això es pot fer a través de capes vectorials Shapefile[16], un estàndard de format desenvolupat per ESRI l'any 1998.

El format Shapefile permet guardar informació no necessàriament relacionada amb la seva topologia, per tant, conjuntament amb els arxius que el Govern del Brasil comparteix a través de la seva web[17] que corresponen als Shapefiles de tots els municipis i estats federals, es pot crear un programa que a partir d'unes coordenades retorni el número d'identificació municipal corresponent, en un seguit de passos.

Figura 15. Shapefile dels estats federals.

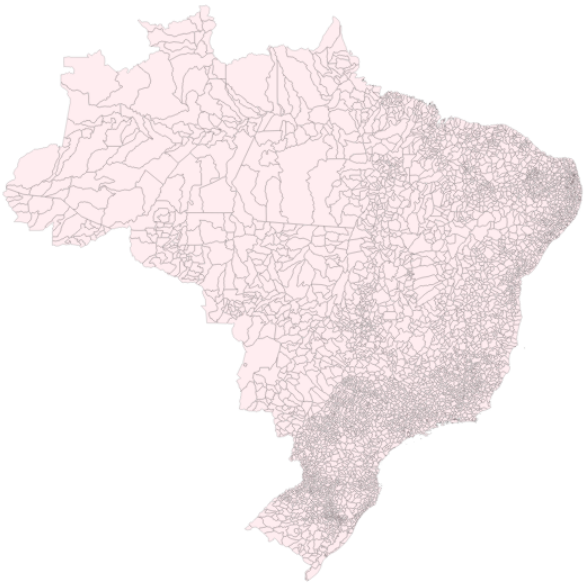

Figura 16. Shapefile dels municipis.

Els passos que es segueixen per dur a terme la georeferenciació inversa són:

1. Trobar en quin estat federal es troben les coordenades introduïdes, si s'escau.
2. Filtrar els municipis de manera que només es considerin els municipis de l'estat federal en qüestió.
3. Buscar en quin municipi dels seleccionats es troben les coordenades.
4. Extreure'n el número identificatiu municipal.

3.2. Primera aproximació: São Paulo

Abans de crear un model generalista que realitzi prediccions per a tots els municipis del Brasil, es va prendre la decisió de crear un model predictiu que englobés una zona més reduïda per tal de comprovar la seva viabilitat amb les dades recollides.

Les variables NDVI, NDWI i temperatura en el format utilitzat tenen una resolució espacial d'uns 5.6 km (o 0.05 graus de diferència entre mostres), per tant es va decidir que la base de dades tingués una forma de malla amb una distància entre punts que coincidís amb aquesta resolució. D'aquesta manera es va seleccionar una zona en forma de quadrat al voltant de São Paulo, zona on cada any hi ha centenars de casos confirmats de dengue concentrats sobretot a l'estiu, és a dir, de mitjans de setembre al març, i principis de tardor, quan hi fa més calor.

Cada costat del quadrat està format per 50 mostres, que conjuntament abarquen una distància d'uns 280 km aproximadament, per tant l'àrea total englobada és d'uns 78.400

km^2. A la figura 17 es mostra una un mapa generat en QGIS que inclou la malla de punts de la base de dades sobre OpenStreetMap.

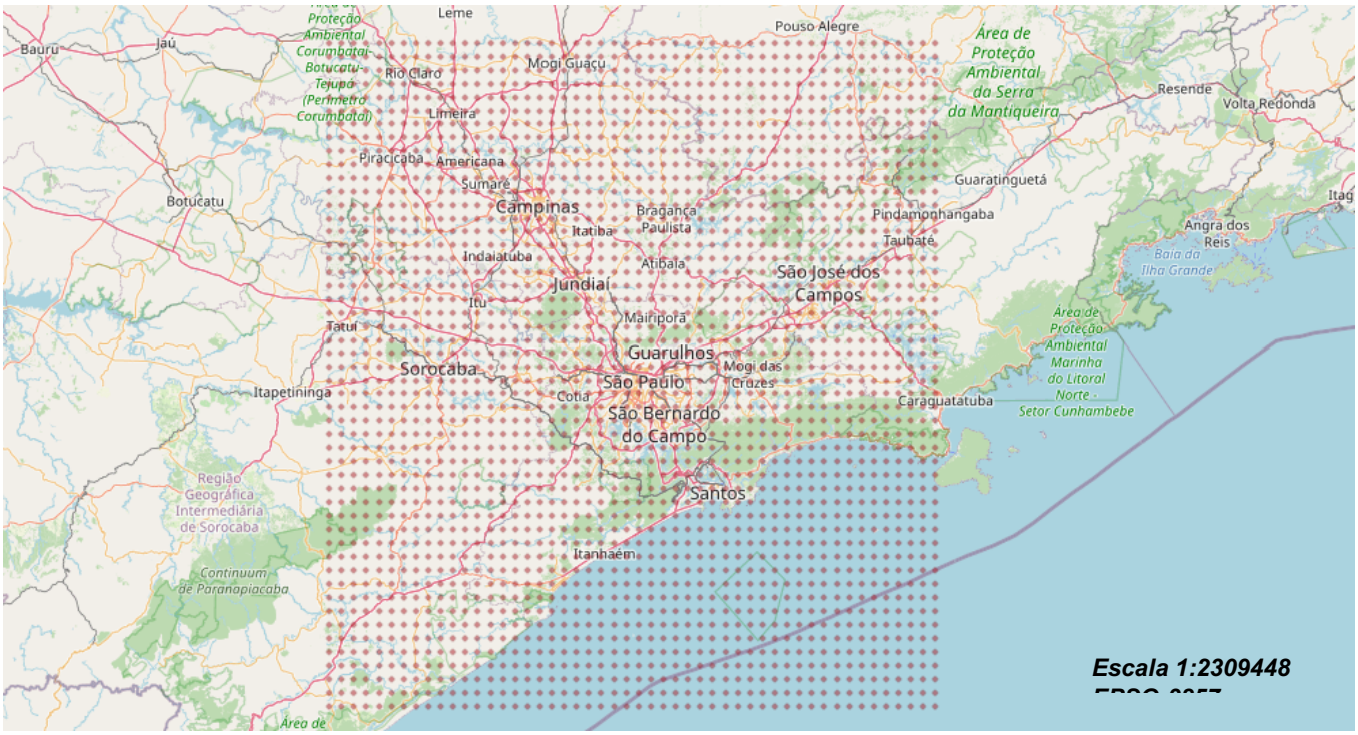

Figura 17. Punts que formen el ràster en forma de malla al voltant de São Paulo.

A cada punt de la malla se l'hi han assignat els valors corresponents de cada variable durant el període de temps que engloba el 2012 i 2013. L'assignació de les variables ambientals s'ha fet mitjançant interpolacions en els casos que ha estat necessari a partir de les dades dels satèl·lits. Per les variables socioeconòmiques s'ha aplicat una georeferenciació inversa a cada punt, assignant així les dades que corresponen al municipi que el conté.

Un cop creada la base de dades *raw*, es descarten el punts que no pertanyen a cap municipi, és a dir, els que corresponen a cossos d'aigua. Després es normalitzen les variables i es poleixen i es transforma la base de dades a format *cooked*, ja a punt per introduir a un model predictiu.

El model que s'ha triat per fer les primeres prediccions ha estat la xarxa neuronal representada a la figura 18. Es basa en una xarxa amb dues *hidden layers* de tres nodes cadascuna. És un sistema *fully connected*, és a dir, que tots els elements d'una capa connecten amb la de la següent.

La xarxa ha estat creada amb el paquet Torch de Pytorch. La seva funció d'activació és la Leaky ReLU en tots els nodes, la funció de pèrdua es calcula amb l'error quadràtic mitjà i fa servir un optimitzador Adam amb un *learning rate* de 0.001.

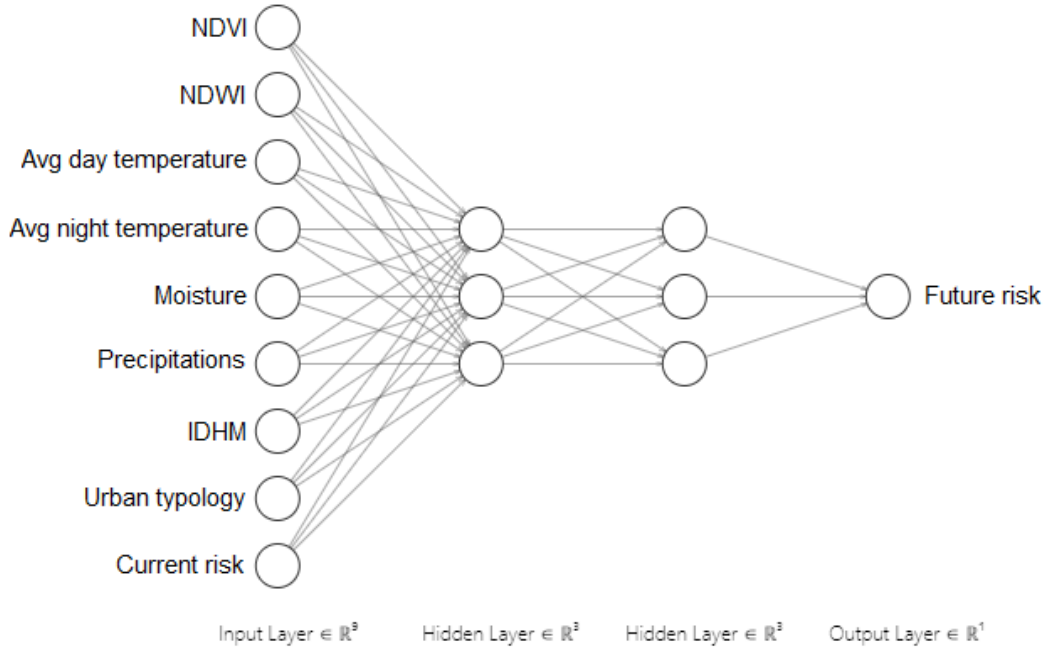

Figura 18. Representació de la xarxa neuronal del primer predictor.

L'entrenament de la xarxa s'ha realitzat amb el 90% de les entrades i el *test* amb el 10% restant, seleccionant quina entrada s'assigna a cada partició de manera completament aleatòria. D'aquesta manera el *training* es fa amb 40.262 mostres i el *test* amb 4.473.

El *batch size*, és a dir, el número de mostres d'entrenament que passen per la xarxa abans d'aplicar el procés d'aprenentatge, és 1. Paral·lelament, en el *training* el número de *epochs* (nombre de passades completes sobre les mostres d'entrenament) és tres. És a dir, en aquest procés totes les entrades es passen tres vegades per la xarxa.

Un cop entrenada la xarxa el següent pas és conèixer la seva fiabilitat. D'aquesta manera s'ha comprovat si les prediccions que es fa sobre les dades de *test* coincideixen amb les dades de risc del següent mes, ja que la diferència de temps entre les dades introduïdes i la predicció generada és d'un mes. S'ha fet servir l'índex de risc A de la taula 6, que descriu cinc possibles estats del risc epidemiològic en cada punt (mínim, baix, mitjà, alt i molt alt). Fent servir aquest índex s'obté que el model prediu de forma correcte el 92.06% de les dades.

Degut a que es podria considerar un índex amb uns marges molt estrets per un model que potencialment té molta variança, si es relaxa el criteri de predicció i es considera com a correcta la predicció de l'estat immediatament proper al real (per exemple si en un punt té un risc "mínim" i es prediu "baix", o si és "molt alt" i es prediu "alt"), la precisió del model s'eleva al 98.28%.

Per tal de visualitzar els resultats s'introdueixen a la xarxa neuronal les dades d'un mes en concret i se n'extreu una predicció. A les figures 19 i 20 es troba la representació del risc epidemiològic a l'abril del 2012 i la seva predicció amb les dades del març.

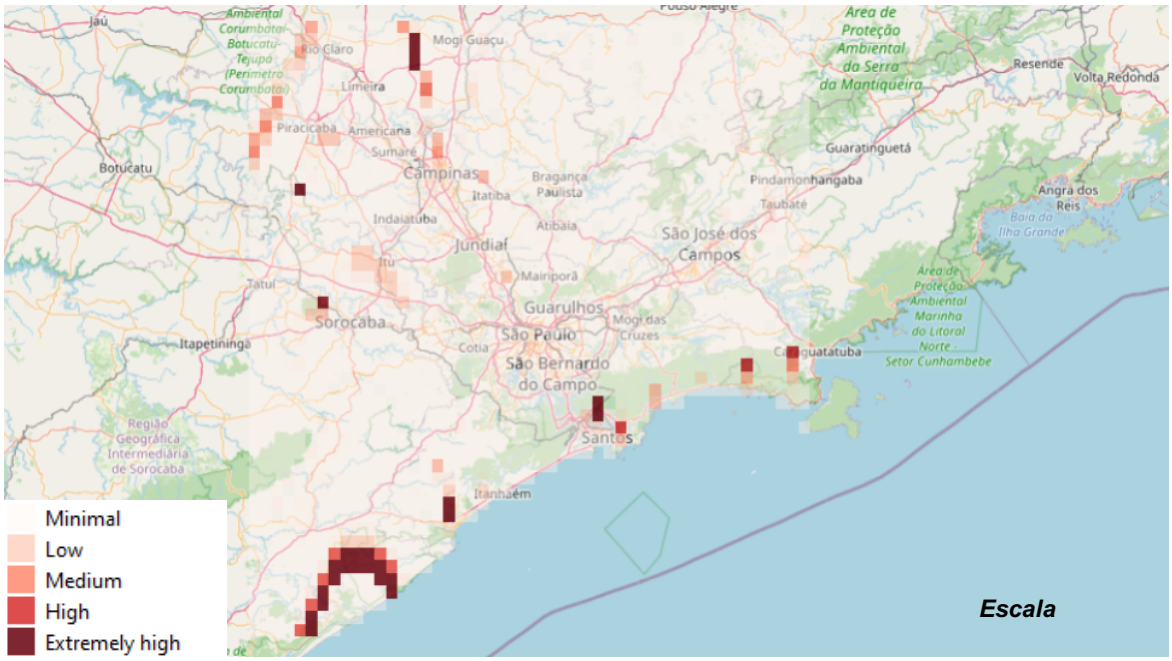

Figura 19. Risc epidemiològic de l'abril del 2012 a l'àrea d'estudi.

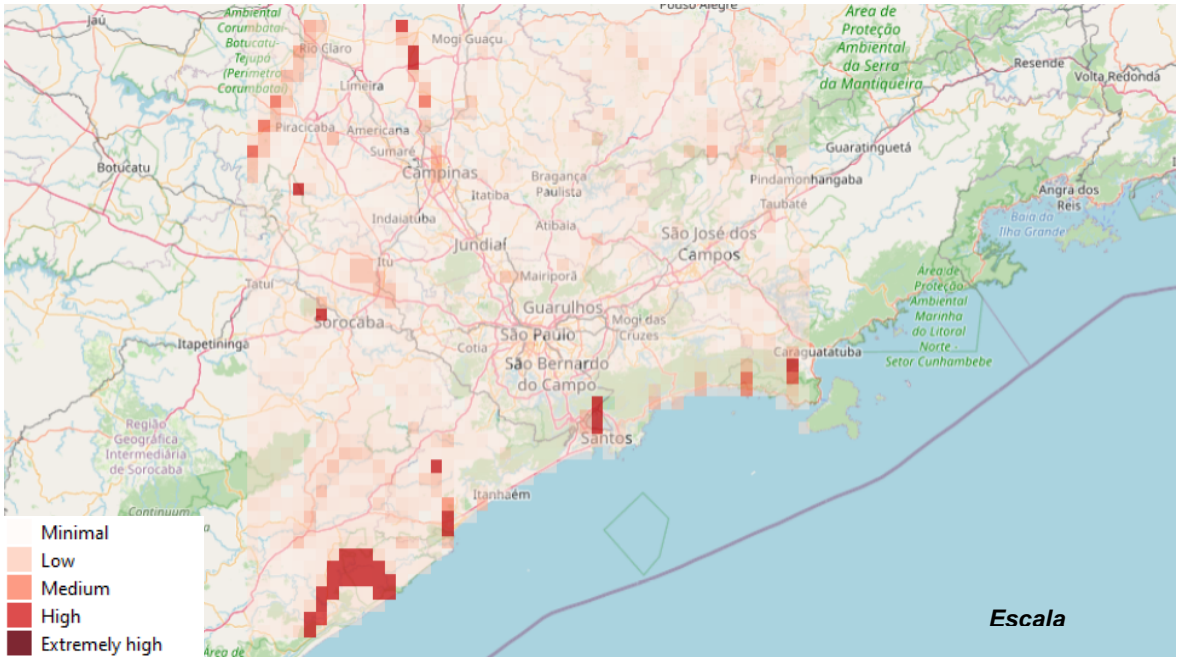

Figura 20. Predicció del risc epidemiològic de l'abril del 2012 a l'àrea d'estudi.

Els resultats són correctes però és legítim pensar que l'alt nivell de predicció és degut a que el risc en la gran majoria de mostres al llarg de tot el període de temps el risc epidemiològic és nul o molt baix. També hi podria haver *overfitting* causat per la redundància generada pel fet que un municipi pugui contenir varis punts dins del ràster,

que tindrien uns valors molt similars. Per aquest motiu en la següent aproximació es segueix un mètode diferent.

3.3. Segona aproximació: prediccions a nivell nacional

Un cop comprovat que és possible generar prediccions amb una diferència de temps entre les mostres i objectiu d'un mes amb un mínim de precisió, el següent pas ha estat crear un model que realitzi un procés similar per tots els municipis del Brasil. Per aquesta finalitat s'ha seguit una metodologia semblant a l'anterior, però amb la principal diferència que les mostres ja no es recullen per punts en forma de malla, sinó que es fa per municipis, tal que cada entrada correspon a un dels 5.564 municipis del Brasil.

Així doncs, es recullen les dades tant ambientals (NDVI, NDWI, humitat,...) com socioeconòmiques (IDHM, tipologia urbana,...) de cadascun dels municipis, realitzant una georeferenciació inversa amb les variables que ho requereixen. Paral·lelament al mètode seguit en la primera aproximació s'ha creat una base de dades *raw* en PostgreSQL que engloba de l'any 2010 al 2013.

3.3.1. Distribució de la base de dades

La base de dades creada és de grans dimensions; concretament la base de dades *cooked* conté 261.367 entrades.

Per crear bons models de predicció és important conèixer bé les dades que es manipulen, així com la seva relació amb la variable a predir. És per això que s'ha intentat visualitzar quina és la relació directa del risc epidemiològic amb les dades del mes anterior fent particions a les dades en varis trams segons els valors de la variable en qüestió, i llavors evaluant com es reparteix el risc del mes següent dins d'aquest tram. Les taules completes que descriuen aquestes particions es troben al *Annex 1*.

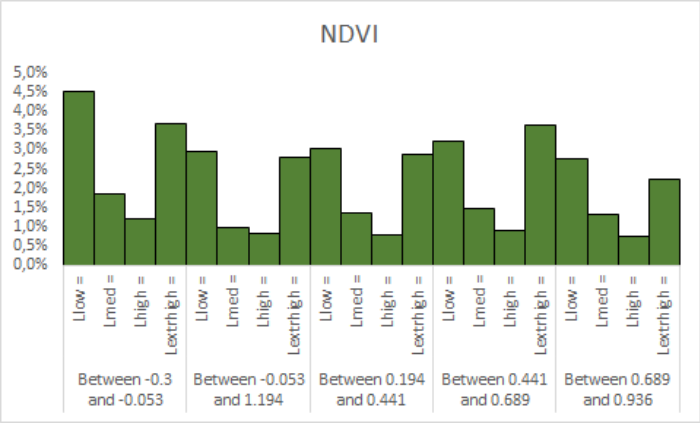

Figura 21. Distribució del risc segons la NDVI del mes anterior.

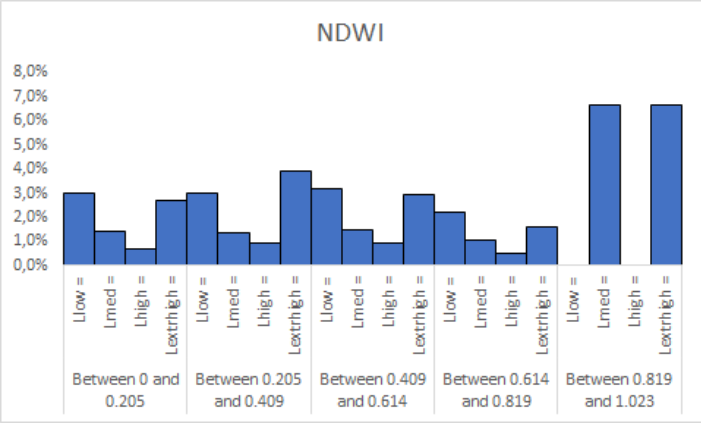

Figura 22. Distribució del risc segons la NDWI del mes anterior.

Com s'observa a les dues figures anteriors, a simple vista no es percep una gran diferència entre el risc epidemiològic segons els diferents trams de l'NDVI i NDWI. Tot i que és cert

que en els valors més elevats de NDWI sí que hi ha un repunt notable en el el risc extremadament elevat, però el número de mostres és massa baix per extreure'n alguna conclusió. Seria lògic doncs, considerar que són variables que per sí soles no serveixen d'indicador, sinó que s'hauria de cercar correlacions amb altres.

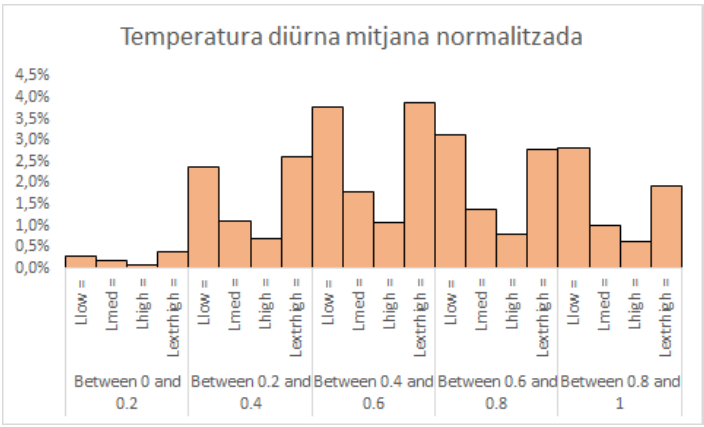

Figura 23. Distribució del risc segons la temperatura diürna del mes anterior.

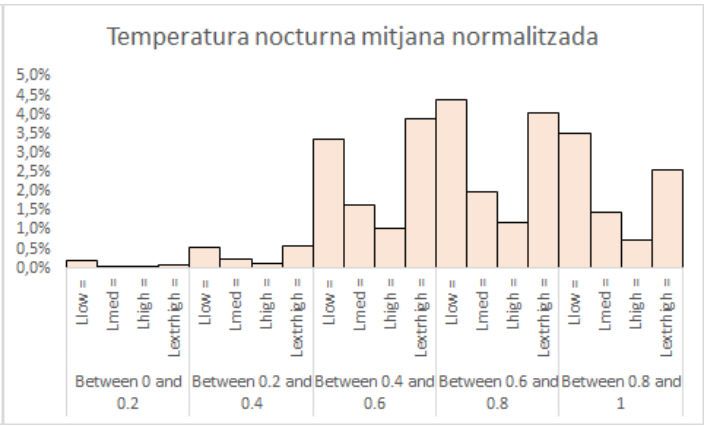

Figura 24. Distribució del risc segons la temperatura nocturna del mes anterior.

En les diferents temperatures sí que s'observa a simple vista com el nombre de casos de dengue es redueix notablement en els seus extrems. L'increment més gran en el risc extremadament alt és del +28,17% en el tercer tram de la temperatura diürna i un +33,15% de la nocturna, mentre que aquest risc decreix un -87,66% i un -97,78% respectivament en el primer tram. És a dir, quan fa més fred el risc és pràcticament nul.

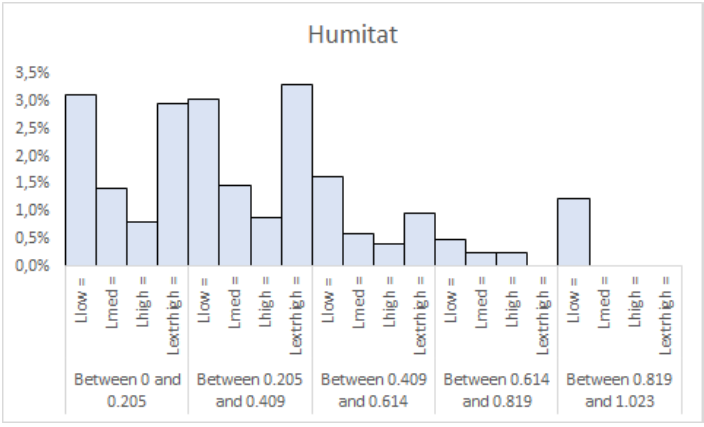

Figura 25. Distribució del risc segons la humitat del mes anterior.

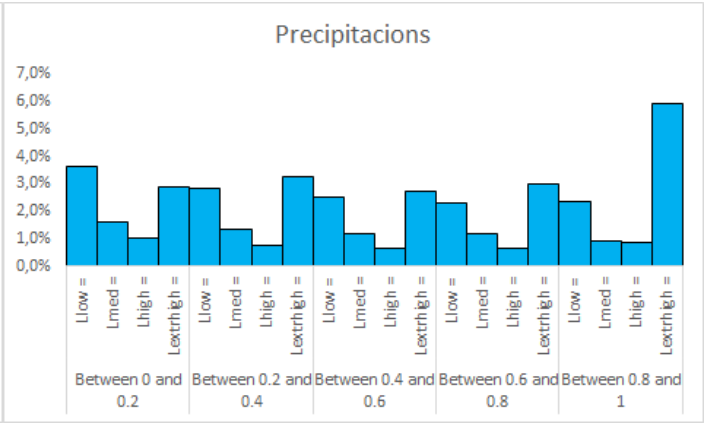

Figura 26. Distribució del risc segons les precipitacions del mes anterior.

La conclusió que es pot extreure segons la distribució de la humitat és que en els punts on aquesta és major, el risc es redueix, però aquest fet es podria atribuir a que aquestes mostres corresponen en la seva majoria a l'Amazones, on el risc ja és més baix de per si.

En canvi, segons la figura 26, quan les precipitacions del mes anterior són extremadament altes sí que es veu un increment clar en el risc més alt, sent aquest un 95,57% superior que el valor global corresponent.

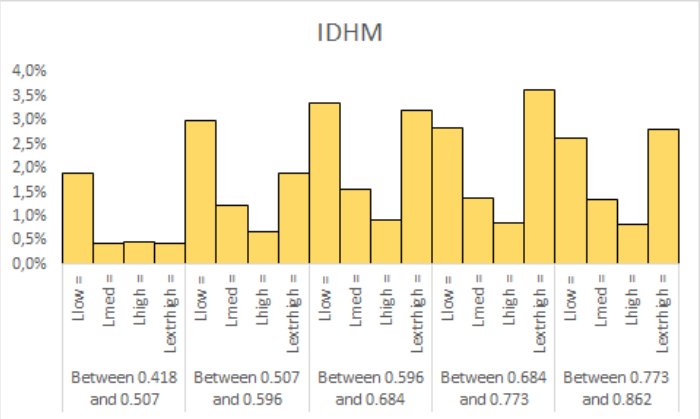

Figura 27. Distribució del risc segons l'IDHM del municipi.

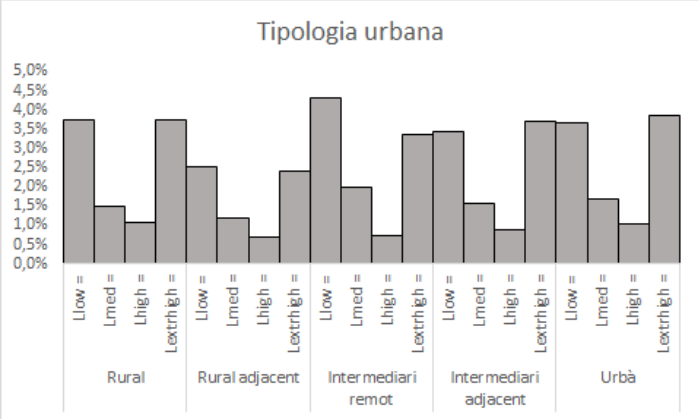

Figura 28. Distribució del risc segons la tipologia urbana del municipi.

En l'IDHM s'observa una disminució de -86,23% en el risc més elevat de dengue en el primer tram, cosa que, com en la humitat, es pot atribuir en part a que una bona representació dels municipis que els conformen es troben a l'Amazones. En canvi en la tipologia urbana no s'observen grans variacions segons els diferents tipus de municipi. Així doncs seria interessant veure com es correlacionaria aquesta variable amb altres, com les precipitacions o l'IDHM.

Un cop arribat a aquest punt ja no és trivial extreure conclusions sobre quin efecte tenen les combinacions d'aquestes variables en el risc epidemiològic. Per aquest motiu, una possible solució és que eines de Machine Learning s'encarreguin d'avaluar-les gràcies a la seva gran capacitat computacional.

3.3.2. Predictor basat en Decision Trees

El model seguit per tal de crear un predictor basat en decision trees ha estat que en lloc de crear-ne un d'únic amb tres possibles estats de sortida, se n'han creat tres per tal de reduir la variança del sistema, tal com es representa a la figura 30.

Cadascun d'aquests arbres és independent i genera una predicció segons la seva finalitat per cada entrada. D'aquesta manera el risc epidemiològic de cada municipi en un moment determinat té tres possibles interpretacions, que en el cas que la predicció fos correcta, l'arbre corresponent marcaria el seu estat actiu mentre els altres dos arbres indicarien "*other*".

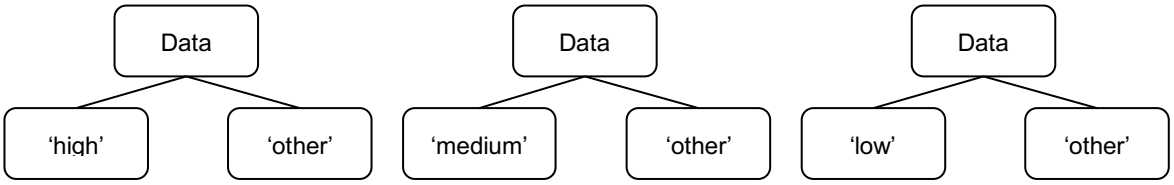

Figura 29. Esquema resum dels tres decision trees utilitzats.

Cada decision tree té una profunditat màxima de 20 capes i es fa servir l'algoritme CART per tal de generar les particions en cada node. Les dades s'han dividit tal que el 70% estan destinades a l'entrenament i el 30% restant al test.

La precisió general dels arbres *high*, *medium* i *low* és 87,81%, 95,21% i 85,02% respectivament, que a primera vista semblen bons números però cal notar que a la gran majoria de mostres el risc és nul o mínim, cosa que pot fer que aquests resultats siguin enganyosos. S'ha comprovat que la precisió de la predicció exacta dels tres estats (excloent estats d'ambigüitats) és del 84,15%, però tenint en compte només els municipis que en el moment de la predicció tenen un risc mitjà o alt, aquesta precisió es redueix al 50,08%, superior al 33% que correspondria a una predicció aleatòria dels tres estats, però no suficientment fiable. La precisió obtinguda és més baixa del que es podria esperar, però cal recordar que la predicció que s'intenta fer és per un objectiu molt ampli, un país sencer, i aquest és un model dels més bàsics.

Per tal de visualitzar una predicció del risc epidemiològic en un mes en concret, primer s'ha generat un mapa amb les dades reals, a la figura 31, del març del 2011.

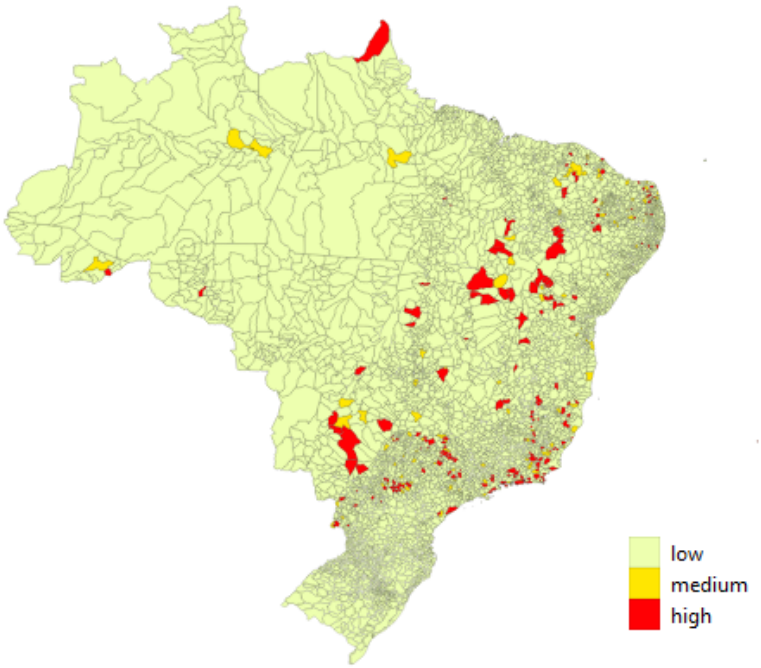

Figura 30. Risc epidemiològic a nivell municipal al març del 2011.

Després per comparar la predicció amb les dades reals s'ha passat pels arbres les dades del febrer del 2011 i se n'ha visualitzat els resultats, a la figura 32.

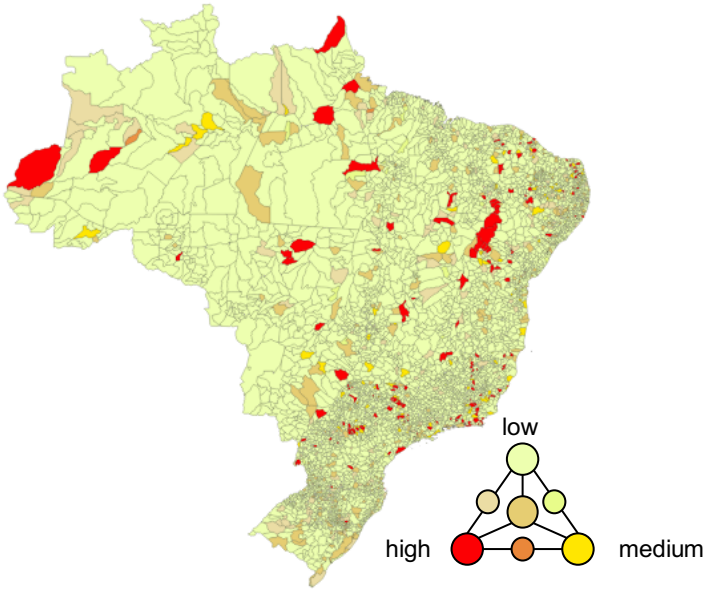

Figura 31. Predicció del risc epidemiològic al març del 2011 amb decision trees i les dades del mes anterior.

Comparant les dues figures s'arriba a la conclusió que la predicció en alguns estats tendeix a ser millor que altres. Per exemple a municipis de l'Amazones es tendeix a predir un risc més elevat que el real, com en aquest cas a Atalaia do Norte i Carauari.

Per tant, es conclou que aquest no és un bon model per generar prediccions del risc epidemiològic a nivell nacional degut possiblement a les peculiaritats de les diferents regions que un model tant esbiaixat com els decision trees no és capaç d'interpretar. Potser podria ser una bona opció si s'encarés el problema des d'una perspectiva més local, com podria ser utilitzar un arbre per un municipi en concret.

3.3.3. Predictor basat en Random Forests

Ja que ha quedat demostrat que els *decision trees* no són un bon model per aquest cas particular. La següent proposta han estat els *random forests*, que com s'ha comentat anteriorment consisteixen en una agrupació de *decision trees* que analitzen particions diferents de les dades.

Inicialment estava planejat que el número d'arbres fos 2.000, però degut a les limitacions de Google Colab, amb 12 Gb màxims de memòria RAM, aquest límit es sobrepassava a l'intentar entrenar un *random forest* d'aquestes dimensions amb la base de dades a nivell nacional. Per aquest motiu, el nombre d'arbres ha estat reduït a 1.500, cosa que podria afectar marginalment la precisió del model.

Les 261.367 entrades de la base de dades s'han dividit tal que un 70% s'ha destinat al *training* i el 30% restant al *test*. Després s'ha aplicat *bootstrapping*, és a dir, que a cada arbre hi arriba un grup de mostres aleatòria d'una mida fixa. En aquest cas en concret el número de mostres que analitza cada arbre equival a un terç del total de mostres de *training*.

A partir del cas d'aplicació anterior, amb *decision trees*, es podria arribar a la conclusió que la propagació del dengue no és igual per cada regió del país tot i tenir unes condicions semblants, i que obviar aquest fet podria ser un dels motius que portin errors en la predicció del risc epidemiològic. És per aquest motiu que s'ha incorporat una nova distinció entre les entrades: l'estat federal on perteneix el municipi. Al ser aquesta una variable classificatòria no-numèrica, s'ha aplicat un criteri de distinció *one-hot* entre els diferents estats per poder-los introduir com a variable al *random forest*, tal com es representa a la taula 10.

state	AC	AL	AM	AP	BA	CE	DF	ES	GO	MA	MG	MS	MT	PA	PB	PE	PI	PR	RJ	RN	RO	RR	RS	SC	SE	SP	TO
CE	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AM	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BA	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MG	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Taula 10. Exemple de codificació one-hot dels estats federals del Brasil.

A diferència de l'aplicació anterior amb *decision trees,* aquest cop s'ha encarat la predicció com una regressió, per posteriorment classificar els resultats segons els diferents índex de risc de la taula 6. Per comprovar el funcionament del model amb *random forests*, s'ha generat una predicció del març del 2013 a partir de les dades del febrer. En aquest cas s'ha obtingut una R^2 de 0.7635, i una precisió del 89.49% a l'avaluar les dades de *test* fent una classificació d'aquestes segons l'índex de risc B.

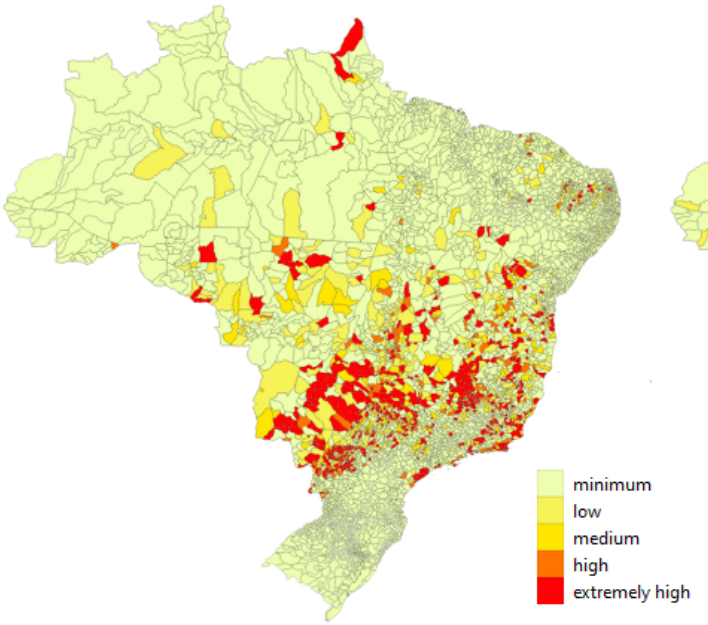

Figura 32. Risc epidemiològic a nivell municipal al març del 2013 segons els casos enregistrats.

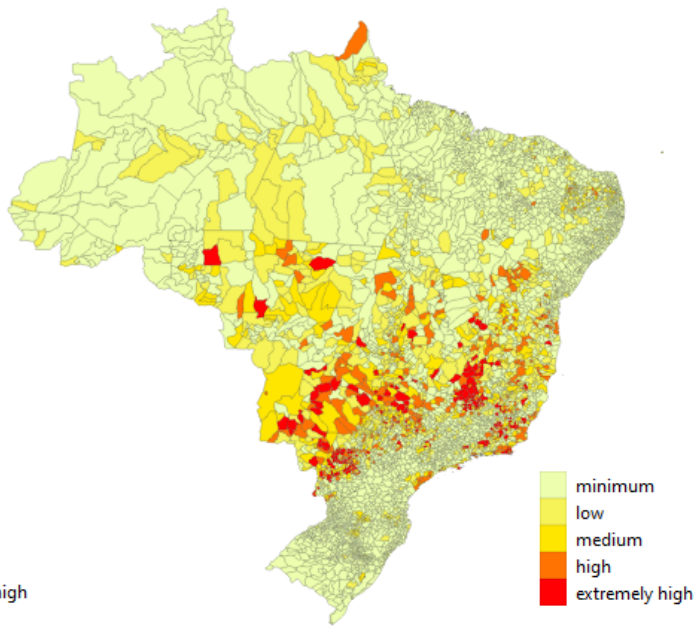

Figura 33. Predicció del risc epidemiològic al març del 2013 amb random forests a partir de les dades del mes anterior.

Com es pot veure aquest model té una precisió superior a l'anterior i és més complicat trobar falsos positius als estats federals amb una incidència menor del dengue. No es detecta algun municipi amb el risc molt alt com del d'Almeirim de l'estat de Parà, però sí que sembla que s'obté una bona precisió en els estats on el dengue té més presència, com Minas Gerais i Mato Grosso do Sul.

3.3.4. Predictor basat en Neural Networks

L'últim model elegit pel desenvolupament d'aquest projecte ha estat una xarxa neuronal que realitzi prediccions pel Brasil en la seva totalitat. Seguint en la línia del predictor anterior s'ha incorporat la variable de l'estat federal de cada municipi codificant-la segons un criteri *one-hot*. En la següent figura es representa l'esquema de la xarxa neuronal en qüestió.

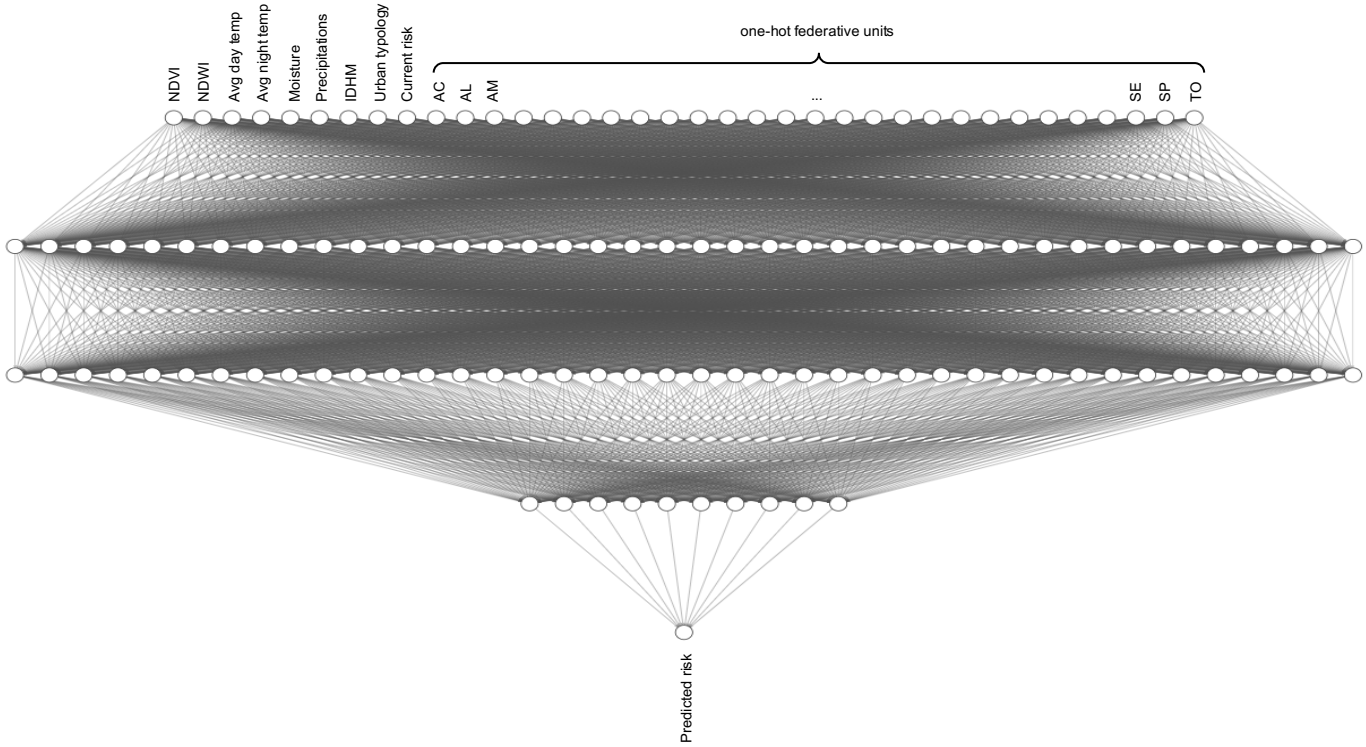

Figura 34. Representació de la xarxa neuronal utilitzada per crear una predicció del risc epidemiològic a nivell nacional.

La arquitectura de la xarxa és de tipus *Deep Feed Forward* (DFF), és a dir, amb més d'una capa oculta. D'aquesta manera l'error es reparteix recursivament d'una forma més extesa, cosa que és positiva per la precisió del model, però que alhora també augmenta el temps d'entrenament d'aquest. La funció de pèrdua triada ha estat la de l'error quadràtic mitjà (fòrmula 13), que es calcula i s'avalua a cada iteració de l'entrenament.

$$mse = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2 \qquad (13)$$

Igual que en el model predictor local de São Paulo s'ha utilitzat un optimitzador Adam, que combina dos mètodes de gradient descendent: el *momentum* i el *Root Mean Square Propagation* (RMSP). El *momentum* es calcula segons les següents expressions:

$$w_{t+1} = w_t - \alpha m_t \qquad (14)$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\left[\frac{\delta L}{\delta w_t}\right] \qquad (15)$$

On "*m*" representa l'agregat dels gradients en un moment determinat. La *"w"* correspon als pesos en aquest mateix moment, la alfa és el *learning rate*, que en aquest cas és 0,001. La *"L"* és la funció de pèrdua (MSE) i beta un paràmetre de moviment mitjà.

Paral·lelament el RMSP es calcula segons:

$$w_{t+1} = w_t - \frac{\alpha_t}{\sqrt{v_t + \epsilon}} * \left[\frac{\delta L}{\delta w_t}\right] \qquad (16)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\delta L}{\delta w_t}\right]^2 \qquad (17)$$

On s'incorpora *"v"*, la suma dels anteriors gradients elevada al quadrat, i epsilon, una petita constant que generalment és 10^{-8}.

El model és *fully-connected* i les funcions d'activació són ReLU. La partició de les entrades s'ha fet tal que un 80% han estat destinades al *training* i el 20% restant al *test*, i l'entrenament s'ha realitzat en cinc *epochs.*

El principal problema ha estat el repartiment de les dades, que tal com es detalla a l'annex 1, en el 90% de les mostres el risc epidemiològic és mínim. Això causa que el model tendeixi a predir riscos molt baixos ja que en moltes àrees i durant molts mesos així es dóna el cas. Per solucionar aquest problema s'han introduït dues condicions prèvies abans de passar una entrada pel model. La primera consisteix en que, si el risc a predir és 0, la mostra tingui un 90% de possibilitats de descartar-se (en la *epoch* d'aquell moment en concret). La segona, aplicada posteriorment, és que si el risc a predir és mínim la mostra

es descartarà amb un 50% de possibilitats. D'aquesta manera queden unes dades més ben repartides per l'entrenament i es tendeix a generar millors prediccions.

Un cop acabat l'entrenament, avaluant la precisió del model segons l'índex de risc B, s'obté una precisió del 92,40%, mentre que si es fa segons l'índex A, aquesta es redueix al 79,21%.

Per veure un exemple d'una predicció obtinguda amb la xarxa neuronal, s'ha fet que aquesta avalui les dades a nivell municipal del gener del 2013, per així poder comparar el risc predit amb el real del febrer del mateix any.

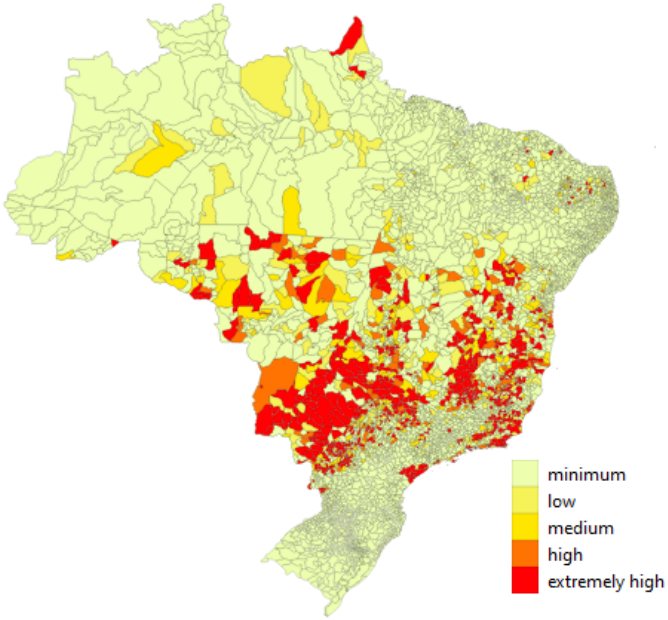

Figura 35. Risc epidemiològic a nivell municipal al febrer del 2013.

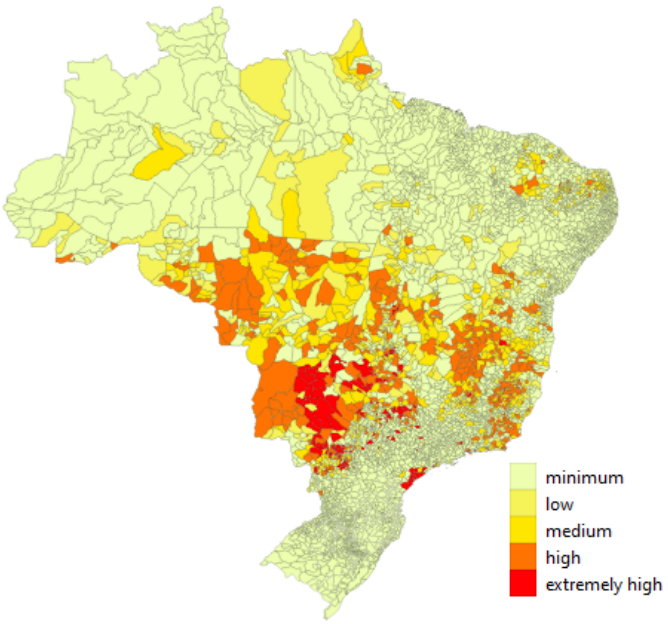

Figura 36. Predicció del risc epidemiològic al febrer del 2013 amb neural networks a partir de les dades del mes anterior.

Com es pot veure, es podria afirmar que el model detecta molts dels municipis on hi hauria un risc important, tot i que en la seva majoria el risc epidemiològic més extrem el detecta com a risc "alt" i no com a "molt alt". Si s'avalua la R^2 d'aquesta predicció comparant-la amb els valors reals s'obté que aquesta és del 0,5041. Per tant es podria dir que és moderadament bona.

4. Resultats i precisió dels models

Un cop completat l'anàlisi amb els diferents models predictius és senzill concloure quins han tingut uns millors resultats, comparant la precisió de les previsions segons els diferents índexs de risc epidemiològic i la R^2 si s'escau.

El millor predictor ha estat clarament el que està basat en *random forests*. Aquest ha obtingut una R^2 de 0.7635 en el cas d'aplicació i una precisió global del 89,49% segons l'índex de risc B.

El següent model que ha obtingut bons resultats ha estat la *neural network* amb una R^2 de 0,5041 en el cas d'aplicació, una precisió del 92,40% segons l'índex de risc epidemiològic B i un 79,21% de l'A.

El predictor que ha obtingut pitjors resultats, cosa que era d'esperar degut a la seva simplicitat, ha estat el basat en *decision trees*. És així que ha obtingut una precisió sense ambigüitats del 85,02%, disminuint al 50,08% si només es té en compte els municipis amb risc real "mitjà" o "alt" segons l'índex B. Cal recordar que en aquest model no s'ha obtingut una R^2 degut a que està encarat com a una classificació directa.

El següent pas per entendre millor els predictors i els factors que afecten més la propagació del dengue podria ser una jerarquització de les variables que s'han fet servir segons la seva importància en la predicció. El paquet de Python Scikit-learn *(sklearn)*, que s'ha utilitzat per a la creació dels *decision trees* i *random forest* en Pytorch, té una prestació que permet extreure la importància de cadascuna de les *features*, o variables, que s'han utilitzat. De manera que a partir d'aquests valors els diferents paràmetres introduïts al model posteriorment es poden ordenar segons aquesta importància en la creació de prediccions.

Els paràmetres més importants dels utilitzats segons els diferents *decision trees* són:

High Risk Decision Tree	Med. Risk Decision Tree	Low Risk Decision Tree
1. *Precipitacions* 2. *Risc anterior* 3. *Temp. diürna* 4. *Temp. nocturna* 5. *Humitat* 6. *NDWI* 7. *NDVI* 8. *IDHM* 9. *Tip. urbana*	1. *Temp. diürna* 2. *Risc anterior* 3. *Temp. nocturna* 4. *Humitat* 5. *Precipitacions* 6. *NDWI* 7. *IDHM* 8. *NDVI* 9. *Tip. urbana*	1. *Risc anterior* 2. *Precipitacions* 3. *Temp. nocturna* 4. *Temp. diürna* 5. *Humitat* 6. *NDWI* 7. *NDVI* 8. *IDHM* 9. *Tip. urbana*

Taula 10. Rànquing d'importància de cada paràmetre en cadascun dels decision trees.

La taula 10 posa en manifest que hi ha algunes petites diferències en l'ordre, però els paràmetres més importants i els que ho són menys sempre són els mateixos. La tipologia urbana de cada municipi sembla que tindria una influència mínima en la predicció, essent aquesta la menys decisiva, seguida de l'IDHM i l'NDVI.

El mateix procediment s'ha seguit pel model basat en *random forests*, que extraient les diferents importàncies de les variables aquestes es poden ordenar. El rànquing de la importància de les variables dins el *random forest* queda de la següent manera:

1. *Risc anterior*
2. *Precipitacions*
3. *Temp. diürna*
4. *Temp. nocturna*
5. *NDWI*
6. *Humitat*
7. *Estat federal*
8. *NDVI*
9. *IDHM*
10. *Tip. urbana*

Com es pot observar les quatre variables que resultarien més importants són: el risc epidemiològic del mes anterior, les precipitacions acumulades i la temperatura diürna i nocturna mitjana. L'estat federal del que forma part cada municipi pel que sembla té una influència clara en la predicció (no és igual la propagació del dengue en les diferents regions del Brasil), però no resulta ser dels paràmetres més determinants.

5. Pressupost

Degut a que les dades necessàries per crear els models de predicció del risc epidemiològic han estat recollides i fetes públiques per tercers, el projecte ha estat centrat únicament en manipulació de dades i software. De la mateixa manera, moltes eines que s'han fet servir (Pyton, Pytorch, Panoply, PostgreSQL, etc) són gratuïtes, i en la major part de programari lliure.

El temps de duració del projecte ha estat de 19 setmanes, per tant les compensacions econòmiques en el seu total s'han calculat segons aquesta xifra. A les següents taules es detallen els costos que sí que es poden atribuir al projecte.

Posició	Persones	Compensació	Dedicació	Total
Enginyer junior	*1*	*12€/h*	*30h setmanals*	*6.840€*
Supervisor	*1*	*25€/h*	*2h setmanals*	*950€*
			Total sous:	**7.790€**

Taula 11. Pressupost relatiu als salaris.

Concepte	Unitats	Preu/unitat	Total
Llicència MATLAB	*1*	*800€/any*	*800€*
Llicència Excel	*1*	*7€/mes*	*28€*
Ordinadors	*2*	*900€*	*1.800€*
		Total material:	**2.628€**

Taula 12. Pressupost relatiu al material.

El pressupost total del projecte correspon a la suma dels dos subtotals anteriors.

Total: 10.418€

6. Conclusions i línies de futur

L'objectiu principal del projecte ha estat assolit: s'han generat algorismes de predicció del dengue per al Brasil utilitzant eines de *Machine Learning* i una conjunció de dades ambientals i socioeconòmiques a nivell municipal, i se n'ha fet una comparativa entre ells. Aquests són un primer resultat i és evident que hi ha marge de millora tant en la recollida de dades ja sigui incrementant la seva cadència, incloent-ne de nous tipus, etc, com en el desenvolupament dels models.

La metodologia que s'ha seguit quedaria resumida en els següents passos:

- Obtenció i preparació de les dades ambientals.
- Obtenció i preparació de les variables socioeconòmiques.
- Recopilació de les dades de casos del dengue a Brasil, per municipis, pel període de l'any 2010 al 2013.
- Creació de la base de dades a partir de totes les dades, fent-les compatibles prèviament.
- Estudi de models de *Machine Learning*.
- Desenvolupament de models de predicció de risc epidemiològic del dengue, primer a la regió de São Paulo i després per tot el Brasil amb diferents models de *Machine Learning*.
- Anàlisi del comportament dels models.

Aquest treball també ha permès aprendre un conjunt d'eines actuals per a l'anàlisi i creació de models predictius i la utilització i visualització de dades georeferenciades, com són el *Machine Learning*, QGIS, Python, Pytorch, Panoply i PostgreSQL.

El predictor que millor ha funcionat ha estat clarament el *random forest*, doncs els altres tendeixen a donar valors baixos en la predicció, i aquest ha estat el que millor ha predit els municipis amb un risc extremadament alt.

Val a dir que aquest és un estudi preliminar de l'aplicació de tècniques de *Machine Learning* en aquest problema en concret i per tant hi ha una sèrie d'actuacions que poden ajudar a millorar aquests models:

- Millora de la resolució espaial de les dades (reduir la distància entre mostres). A més es podria analitzar quina és la millor tècnica d'interpolació per tenir totes les dades amb la mateixa referència.
- Estudiar l'efecte de la resolució temporal. En aquest projecte s'ha treballat amb dades mensuals però seria interessant estudiar l'efecte de treballar amb una periodicitat quinzenal o inferior ja que l'*Aedes Aegypti*, tal com s'ha exposat a la introducció del document, té un cicle reproductiu d'uns 10-15 dies, cosa que fa que una freqüència d'un mes entre mostres es pugui considerar insuficient.
- Estudiar la incorporació d'altres variables a la base de dades. Així com analitzar l'ús de diferents estadístiques obtingudes a partir de les variables, com el valor màxim, mínim, desviació estàndard o anomalies.

Per evitar redundàncies el millor mètode ha estat el de fer servir tant les variables ambientals com socioeconòmiques en format municipal realitzant una georeferenciació

inversa en les que fos necessari. Però per obtenir millors resultats el següent pas seria repensar com s'extreuen els múltiples valors dels punts en forma de malla que conformen les dades de satèl·lits dins d'un mateix terme municipal, i així poder fer un millor anàlisi amb aquestes dades.

Per altra banda, si es mantingués el format de dades mixt (dades ambientals en ràster en forma de malla i les dades socioeconòmiques i de casos de dengue en format municipal), seria possible entrenar una xarxa neuronal amb capes convolucionals. Les dades d'un període de temps en concret tindrien una relació espaial obvia entre elles. Per tant, es podrien fer convolucions en 2D i fins i tot 3D si s'incorporés el temps de la mostra en el model. D'aquesta manera al poder relacionar un punt de la malla amb els més propers segurament faria que la distinció de l'estat federal fos innecessària i la predicció més precisa, ja que hi ha molts motius per considerar que si el risc en un punt és elevat, hi ha més possibilitats que a un punt immediatament proper també ho acabi sent.

Bibliografia

[1] "Dengue". [Online] Available: https://ca.wikipedia.org/wiki/Dengue. [Accessed: 23 November 2021].

[2] Manuel Espinosa, Eliana Marina Alvarez Di Fino, Marcelo Abril,Mario Lanfri, Maria Victoria Periago, Carlos Marcelo Scavuzzo. "Operational satellite-based temporal modelling of Aedes population in Argentina". *4 September 2018. Buenos Aires, Argentina. Geospatial Health 2018; volume 13:734*

[3] "Mosquito life cycle". [Online] Available: https://www.baycounty-mi.gov/MosquitoControl/MosquitoLifeCycle.aspx

[4] Anh Kim Nguyen, Yuei-An Liou. "An approach for risk maps of vector-borne infectious diseases: ecological and adaptive capacity indicators". *Center for Space and Remote Sensing Research, National Central University, Taiwan Institute of Geography. Vietnam Academy of Science and Technology, Hanoi Viet Nam. Taiwan Group on Earth Observations, Hsinchu, Taiwan, R.O.C. 978-1-5386-7150-4/18/$31.00 ©2018 IEEE*

[5] "MODIS: Moderate Resolution Imaging Spectroradiometer". [Online] Available: https://modis.gsfc.nasa.gov

[6] Bo-cai Gao. "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space". *Remote Sensing of Environment, Volume 58, Issue 3, December 1996.*

[7] S. K. McFeeters (1996) "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features", *International Journal of Remote Sensing, 17:7, 1425-1432, DOI: 10.1080/01431169608948714.*

[8] "LP DAAC: The Land Processes Distributed Active Archive Center" [Online] Available: https://e4ftl01.cr.usgs.gov/MOLT/MOD13C2.006/

[9] "Soil Moisture". [Online] Available: https://www.drought.gov/drought/data-maps-tools/soil-moisture

[10]Yusselfy Márquez Benítez, Katherine Johana Monroy Cortés, Edna Geraldine Martínez Montenegro, Víctor Hugo Peña García, Ángela Liliana Monroy Díaz. "Influence of environmental temperature in the mosquito Aedes and the transmission of the dengue virus". *2018 CES Medicina. SSN 0120-8705*

[11]Cristina Parada. "A collection of 79 attributes from Brazilian Citie" [Online] Available: https://www.kaggle.com/crisparada/brazilian-cities

[12]Rachel Lowe, Christovam Barcellos, Caio A S Coelho, Trevor Bailey, Giovannini Evelim Coelho, Richard Graham, Tim Jupp, Walter Massa Ramalho, Marilia Sá Carvalho, David B Stephenson, Xavier Rodó. "Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts". *Lancet Infect Dis 2014. 14:619-26*

[13]"Human Development Reports". [Online] Available: http://hdr.undp.org/en/reports/global/2004/espanol

[14]"Backpropagation". [Online] Available: https://en.wikipedia.org/wiki/Backpropagation

[15] Andreas Hackeloeer, Klaas Klasing, Jukka M. Krisp & Liqiu Meng (2014) "Georeferencing: a review of methods and applications", *Annals of GIS, 20:1, 61-69, DOI: 10.1080/19475683.2013.868826*

[16] "ESRI Shapefile Technical Description". *July 1998 An ESRI White Paper*

[17] "Municipal mesh database". [Online] Available: https://www.ibge.gov.br/en/geosciences/full-list-geosciences/18890-municipal-mesh.html

Annex 1: distribució del risc epidemiològic a nivell municipal a partir de les dades del mes anterior

GLOBAL		
		Lmin = 91,7048%
		Llow = 3,0306%
Count =	261367	Lmed = 1,3988%
		Lhigh = 0,8306%
		Lextrhigh = 3,0352%

Taula 13. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent.

NDVI		%dif	
Between -0.3 and -0.053	Lmin = 88,7868%	-3,18%	Count = 1088
	Llow = 4,5037%	48,61%	
	Lmed = 1,8382%	31,41%	
	Lhigh = 1,1949%	43,86%	
	Lextrhigh = 3,6765%	21,13%	
Between -0.053 and 1.194	Lmin = 92,3967%	0,75%	Count = 605
	Llow = 2,9752%	-1,83%	
	Lmed = 0,9917%	-29,10%	
	Lhigh = 0,8264%	-0,51%	
	Lextrhigh = 2,8099%	-7,42%	
Between 0.194 and 0.441	Lmin = 91,9444%	0,26%	Count = 39724
	Llow = 3,0385%	0,26%	
	Lmed = 1,3594%	-2,82%	
	Lhigh = 0,7930%	-4,53%	
	Lextrhigh = 2,8648%	-5,61%	
Between 0.441 and 0.689	Lmin = 90,7752%	-1,01%	Count = 128328
	Llow = 3,2074%	5,83%	
	Lmed = 1,4564%	4,12%	
	Lhigh = 0,9047%	8,92%	
	Lextrhigh = 3,6563%	20,46%	
Between 0.689 and 0.936	Lmin = 92,9329%	1,34%	Count = 91622
	Llow = 2,7624%	-8,85%	
	Lmed = 1,3326%	-4,73%	
	Lhigh = 0,7389%	-11,04%	
	Lextrhigh = 2,2331%	-26,43%	

Taula 14. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de NDVI.

NDWI

Between 0 and 0.205	Lmin = 92,1934%	0,53%	Count = 38726
	Llow = 3,0004%	-1,00%	
	Lmed = 1,4189%	1,44%	
	Lhigh = 0,6837%	-17,69%	
	Lextrhigh = 2,7037%	-10,92%	
Between 0.205 and 0.409	Lmin = 90,7784%	-1,01%	Count = 69023
	Llow = 3,0019%	-0,95%	
	Lmed = 1,3821%	-1,19%	
	Lhigh = 0,9026%	8,67%	
	Lextrhigh = 3,9349%	29,64%	
Between 0.409 and 0.614	Lmin = 91,5116%	-0,21%	Count = 112542
	Llow = 3,1961%	5,46%	
	Lmed = 1,4732%	5,32%	
	Lhigh = 0,8966%	7,95%	
	Lextrhigh = 2,9225%	-3,71%	
Between 0.614 and 0.819	Lmin = 94,6000%	3,16%	Count = 32222
	Llow = 2,2252%	-26,58%	
	Lmed = 1,0676%	-23,68%	
	Lhigh = 0,5276%	-36,48%	
	Lextrhigh = 1,5797%	-47,95%	
Between 0.819 and 1.023	Lmin = 86,6667%	-5,49%	Count = 15
	Llow = 0,0000%	-100,00%	
	Lmed = 6,6667%	376,60%	
	Lhigh = 0,0000%	-100,00%	
	Lextrhigh = 6,6667%	119,65%	

Taula 15. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de NDWI.

Temperatura diürna mitjana normalitzada

Between 0 and 0.2	Lmin = 99,0746%	8,04%	Count =	13616
	Llow = 0,2864%	-90,55%		
	Lmed = 0,1910%	-86,35%		
	Lhigh = 0,0734%	-91,16%		
	Lextrhigh = 0,3746%	-87,66%		
Between 0.2 and 0.4	Lmin = 93,2088%	1,64%	Count =	70488
	Llow = 2,3777%	-21,54%		
	Lmed = 1,1023%	-21,20%		
	Lhigh = 0,6824%	-17,84%		
	Lextrhigh = 2,6288%	-13,39%		
Between 0.4 and 0.6	Lmin = 89,4786%	-2,43%	Count =	107913
	Llow = 3,7725%	24,48%		
	Lmed = 1,7950%	28,32%		
	Lhigh = 1,0638%	28,08%		
	Lextrhigh = 3,8902%	28,17%		
Between 0.6 and 0.8	Lmin = 91,9066%	0,22%	Count =	57578
	Llow = 3,1297%	3,27%		
	Lmed = 1,3842%	-1,04%		
	Lhigh = 0,7954%	-4,24%		
	Lextrhigh = 2,7840%	-8,28%		
Between 0.8 and 1	Lmin = 93,5950%	2,06%	Count =	11772
	Llow = 2,8287%	-6,66%		
	Lmed = 1,0109%	-27,73%		
	Lhigh = 0,6286%	-24,32%		
	Lextrhigh = 1,9368%	-36,19%		

Taula 16. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de la temperatura diürna mitjana.

Temperatura nocturna mitjana normalitzada

Between 0 and 0.2	Lmin = 99,6633%	8,68%	Count = 8910
	Llow = 0,1908%	-93,70%	
	Lmed = 0,0561%	-95,99%	
	Lhigh = 0,0224%	-97,30%	
	Lextrhigh = 0,0673%	-97,78%	
Between 0.2 and 0.4	Lmin = 98,5539%	7,47%	Count = 53247
	Llow = 0,5296%	-82,52%	
	Lmed = 0,2272%	-83,76%	
	Lhigh = 0,1164%	-85,99%	
	Lextrhigh = 0,5728%	-81,13%	
Between 0.4 and 0.6	Lmin = 90,0809%	-1,77%	Count = 92186
	Llow = 3,3519%	10,60%	
	Lmed = 1,6402%	17,26%	
	Lhigh = 1,0240%	23,28%	
	Lextrhigh = 3,9030%	28,59%	
Between 0.6 and 0.8	Lmin = 88,3932%	-3,61%	Count = 86527
	Llow = 4,4032%	45,29%	
	Lmed = 1,9901%	42,27%	
	Lhigh = 1,1719%	41,09%	
	Lextrhigh = 4,0415%	33,15%	
Between 0.8 and 1	Lmin = 91,7346%	0,03%	Count = 20495
	Llow = 3,5228%	16,24%	
	Lmed = 1,4443%	3,25%	
	Lhigh = 0,7270%	-12,47%	
	Lextrhigh = 2,5714%	-15,28%	

Taula 17. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de la temperatura nocturna mitjana.

Humitat

Between 0 and 0.162	Lmin = 91,6936%	-0,01%	Count = 103510
	Llow = 3,1282%	3,22%	
	Lmed = 1,4057%	0,49%	
	Lhigh = 0,8028%	-3,35%	
	Lextrhigh = 2,9698%	-2,15%	
Between 0.162 and 0.323	Lmin = 91,2946%	-0,45%	Count = 135330
	Llow = 3,0584%	0,92%	
	Lmed = 1,4720%	5,23%	
	Lhigh = 0,8815%	6,13%	
	Lextrhigh = 3,3082%	8,99%	
Between 0.323 and 0.485	Lmin = 96,3979%	5,12%	Count = 9328
	Llow = 1,6402%	-45,88%	
	Lmed = 0,5896%	-57,85%	
	Lhigh = 0,4074%	-50,95%	
	Lextrhigh = 0,9648%	-68,21%	
Between 0.485 and 0.647	Lmin = 99,0419%	8,00%	Count = 835
	Llow = 0,4790%	-84,19%	
	Lmed = 0,2395%	-82,88%	
	Lhigh = 0,2395%	-71,17%	
	Lextrhigh = 0,0000%	-100,00%	
Between 0.647 and 0.808	Lmin = 98,7805%	7,72%	Count = 82
	Llow = 1,2195%	-59,76%	
	Lmed = 0,0000%	-100,00%	
	Lhigh = 0,0000%	-100,00%	
	Lextrhigh = 0,0000%	-100,00%	

Taula 18. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de la humitat.

Precipitacions

Between 0 and 0.2	Lmin = 90,8648%	-0,92%	Count = 97535
	Llow = 3,6397%	20,10%	
	Lmed = 1,6261%	16,25%	
	Lhigh = 0,9966%	19,99%	
	Lextrhigh = 2,8728%	-5,35%	
Between 0.2 and 0.4	Lmin = 91,7859%	0,09%	Count = 89967
	Llow = 2,8410%	-6,26%	
	Lmed = 1,3316%	-4,80%	
	Lhigh = 0,7781%	-6,32%	
	Lextrhigh = 3,2634%	7,52%	
Between 0.4 and 0.6	Lmin = 92,8665%	1,27%	Count = 51055
	Llow = 2,5149%	-17,02%	
	Lmed = 1,2085%	-13,60%	
	Lhigh = 0,6659%	-19,83%	
	Lextrhigh = 2,7441%	-9,59%	
Between 0.6 and 0.8	Lmin = 92,8726%	1,27%	Count = 18969
	Llow = 2,3248%	-23,29%	
	Lmed = 1,1598%	-17,09%	
	Lhigh = 0,6590%	-20,66%	
	Lextrhigh = 2,9838%	-1,69%	
Between 0.8 and 1	Lmin = 89,9245%	-1,94%	Count = 3841
	Llow = 2,3431%	-22,69%	
	Lmed = 0,9112%	-34,86%	
	Lhigh = 0,8852%	6,57%	
	Lextrhigh = 5,9360%	95,57%	

Taula 19. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de les precipitacions acumulades.

IDHM			
Between 0.418 and 0.507	Lmin = 96,8085%	5,57%	Count = 2632
	Llow = 1,8997%	-37,32%	
	Lmed = 0,4179%	-70,12%	
	Lhigh = 0,4559%	-45,11%	
	Lextrhigh = 0,4179%	-86,23%	
Between 0.507 and 0.596	Lmin = 93,1912%	1,62%	Count = 58938
	Llow = 2,9930%	-1,24%	
	Lmed = 1,2352%	-11,70%	
	Lhigh = 0,6719%	-19,11%	
	Lextrhigh = 1,9088%	-37,11%	
Between 0.596 and 0.684	Lmin = 90,9345%	-0,84%	Count = 90287
	Llow = 3,3515%	10,59%	
	Lmed = 1,5750%	12,60%	
	Lhigh = 0,9182%	10,55%	
	Lextrhigh = 3,2208%	6,11%	
Between 0.684 and 0.773	Lmin = 91,3288%	-0,41%	Count = 100298
	Llow = 2,8316%	-6,57%	
	Lmed = 1,3659%	-2,35%	
	Lhigh = 0,8545%	2,88%	
	Lextrhigh = 3,6192%	19,24%	
Between 0.773 and 0.862	Lmin = 92,3795%	0,74%	Count = 9212
	Llow = 2,6162%	-13,67%	
	Lmed = 1,3569%	-3,00%	
	Lhigh = 0,8359%	0,64%	
	Lextrhigh = 2,8116%	-7,37%	

Taula 20. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons el valor de IDHM.

Tipologia urbana

Rural	Lmin = 89,9593%	-1,90%	Count =	15228
	Llow = 3,7234%	22,86%		
	Lmed = 1,4972%	7,03%		
	Lhigh = 1,0835%	30,45%		
	Lextrhigh = 3,7365%	23,11%		
Rural adjacent	Lmin = 93,1681%	1,60%	Count =	142786
	Llow = 2,5367%	-16,30%		
	Lmed = 1,1899%	-14,93%		
	Lhigh = 0,6835%	-17,71%		
	Lextrhigh = 2,4218%	-20,21%		
Intermediari remot	Lmin = 89,5745%	-2,32%	Count =	2820
	Llow = 4,3262%	42,75%		
	Lmed = 1,9858%	41,96%		
	Lhigh = 0,7447%	-10,34%		
	Lextrhigh = 3,3688%	10,99%		
Intermediari adjacent	Lmin = 90,3789%	-1,45%	Count =	32148
	Llow = 3,4434%	13,62%		
	Lmed = 1,5864%	13,41%		
	Lhigh = 0,8990%	8,24%		
	Lextrhigh = 3,6923%	21,65%		
Urbà	Lmin = 89,7492%	-2,13%	Count =	68385
	Llow = 3,6602%	20,77%		
	Lmed = 1,7007%	21,58%		
	Lhigh = 1,0529%	26,76%		
	Lextrhigh = 3,8371%	26,42%		

Taula 21. Distibució del nombre d'entrades a la base de dades segons el risc epidemiològic del mes següent discriminant segons la tipologia urbana del municipi.

Glossari

NDVI: Normalized difference vegetation index

NDWI: Normalized difference water index

MODIS: Moderate Resolution Imaging Spectroradiometer

SMOS: Soil Moisture and Ocean Salinity satellite

NASA: National Aeronautics and Space Administration

NIR: Near-infrared

VIS: Visible

MIR: Mid-infrared

LP DAAC: The Land Processes Distributed Active Archive Center

TRMM: Tropical Rainfall Measuring Mission

GPM: Global Precipitation Measurement

EV: Esperança de Vida

IAA: Índex d'Alfabetització Adulta

TCA: Taxa Combinada d'Alfabetització

PIBpc: PIB per capita tenint en compte la PPA en $ EUA

CART: Classification And Regression Trees